UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA ANIMAL

Patterns of positive selection on the transcriptome of western Iberian *Squalius* fish: a new approach accounting for alternative splicing

Carlos Ernesto Rodríguez Ramírez

Mestrado em Biologia Evolutiva e do Desenvolvimento

Dissertação orientada por:
Professora Doutora Maria Manuela Coelho
Professor Doutor Vítor Sousa

2019

For my family,

who no matter the distance always supported me.

Acknowledgments (Agradecimentos)

Como é sempre o caso em todas as nossas conquistas importantes na vida, este trabalho não teria sido de todo possível sem a ajuda e o apoio de muita gente. Estou-vos profundamente grato a todos vocês, não só pelo vosso apoio durante a realização deste trabalho, mas pelo apoio incondicional que sempre me deram. Palavras nunca serão suficientes para expressar a gratidão que sinto. Ainda assim, deixo-vos um breve, porém sentida, mensagem de agradecimento a todos vocês.

Primeiro quero agradecer à minha amada família. É graças ao vosso grande apoio e amor que hoje sou a pessoa que sou e que consegui chegar até onde cheguei. Mesmo agora, a um oceano de distância, vocês continuam a apoiar-me incondicionalmente. Apesar de apenas vos dedicar uma fração do tempo que merecem, vocês continuam a arranjar tempo para mim sempre que necessito. Pelo amor e apoio que me deram a vida inteira, obrigado. Amo-vos muito.

Em segundo lugar, queria agradecer aos meus fantásticos orientadores: a Professora Manuela Coelho e o Vítor Sousa. Obrigado pelo vosso constante apoio, positivismo, energia e paciência. Obrigado por todas as fascinantes discussões científicas que tivemos (eram sempre o ponto alto do meu dia). Obrigado por tudo o que me ensinaram. Tenho a certeza que hoje sou um melhor investigador graças a vocês e estou muito agradecido por ter tido a oportunidade de trabalhar com vocês.

A seguir, queria agradecer as pessoas fantásticas do meu grupo Evolutionary Genetics. Foram vocês que animaram o meu dia-a-dia durante os longos meses da realização deste trabalho. Obrigado por todas as discussões interessantes que tivemos (tanto as científicas como as menos científicas!). Obrigado por toda o apoio, ajuda, sugestões e conselhos que me deram. Fico feliz por fazermos todos parte do mesmo grupo.

Quero agradecer ao João Moreno pela sua inestimável ajuda na parte da extração do *RNA* do *Squalius aradensis* e do *Squalius torgalensis* e no envio das amostras para a BGI. Foi um processo mais sofrido do que estávamos à espera, mas nunca desististe de repeti-lo até as amostras chegarem em boas condições à BGI. Também te agradeço por graciosamente teres cedido acesso às tuas sequências de Sanger de genes do ciclo circadiano que foram essenciais para demonstrar a fragmentação das sequências do transcriptoma.

Tamém quero agradecer ao Miguel Machado e ao Tiago Jesus. Ao Miguel Machado por graciosamente ter-nos dado acesso aos dados de *RNA-seq* não-publicados do *Squalius alburnoides* (AA) e ao Tiago por ter partilhado conosco múltiplos scripts para análises transcriptomicas. Para além disso, ambos tiveram a simpatia de explicar-nos a organização do servidor do Evolutionary Genetics.

E claro que não podia deixar de agradecer a todos meus maravilhosos amigos. Apesar de já não nós vermos com a frequência que gostaríamos, sempre que nos reencontramos é motivo de grande alegria para mim.

São vocês que me levantam o animo e põem um sorriso na cara independentemente de quão desanimado esteja. Infelizmente não vos posso agradecer a todos individualmente aqui, mas há duas pessoas em particular a quem tenho que agradecer. Primeiro, Catarina Mouta, por vir visitar-me a faculdade sempre que podia. Não tens ideia do quão valiosas estas visitas eram para mim: pela companhia, apoio e alegria que me davas. Obrigado, Cata. Segundo, Inês Ventura, por todo o apoio e bons conselhos que me deste, e por me levantares sempre o ânimo com a tua energia contagiante. Obrigado, amiga.

Finalmente, quero agradecer a minha namorada, Cátia Chanfana. Estiveste sempre ao meu lado. Festejaste comigo os pontos altos, e sofreste comigo os pontos baixos O teu amor, carinho e apoio incondicional são constantes que alegraram os meus dias, mesmo os mais difíceis, e que me deram força para ultrapassar os obstáculos que tinha pela frente. És uma pessoa maravilhosa e tenho muita sorte de que faças parte da minha vida. Estou-te profundamente grato por tudo, meu amor. Obrigado.

Resumo

Um dos principais objetivos em biologia evolutiva é compreender os mecanismos moleculares do processo de adaptação dos organismos vivos, nomeadamente o tempo e o modo em que este ocorre. Avanços na sequenciação de última geração ("next generation sequencing") têm permitido melhorar o nosso conhecimento sobre os mecanismos da adaptação ao nível molecular, inclusive em espécies não-modelo. Um exemplo disto é o uso de dados de *RNA-seq* para procurar assinaturas de seleção positiva ao nível do transcriptoma usando o rácio entre substituições não-sinónimas e substituições sinónimas (dN/dS). Contudo, a utilização de dados de *RNA-seq* nestes estudos está associado ao risco de existirem diferentes isoformas resultantes de *splicing* alternativo misturadas nos alinhamentos de genes ortologos. Ainda assim, ao providenciar dezenas de milhares de sequências de genes codificantes, os dados de *RNA-seq* podem ajudar a compreender o tempo e o modo do processo adaptativo. Actualmente, também existem métodos para inferir o rácio dN/dS para cada ramo de uma árvore de genes ("*gene tree")*, denominados modelos de *branch-site*. Estes métodos permitem identificar os ramos das "*gene trees"* onde é mais provável ter ocorrido seleção positiva para determinado gene.

Em Portugal existem quatro espécies de ciprinídeos de água doce do género *Squalius* (*S. carolitertii, S. pyrenaicus, S. torgalensis* e *S. aradensis*) que estão distribuídas ao longo de um *cline* de temperatura norte-sul que abrange dois tipos de climas – Atlântico e Mediterrânico - e que é refletido num cline de características morfológicas. Estudos recentes de expressão génica e previsão da estrutura das proteínas em duas destas espécies, *S. torgalensis* e *S. carolitertii,* sugeriram que *S. torgalensis* possui uma adaptação às temperaturas mais elevadas, características da região Mediterrânica que habita, nomeadamente no gene *hsp90*.

Neste estudo procurámos determinar se existem evidencias de seleção positiva no transcriptoma de quatro espécies de *Squalius* do oeste da Península Ibérica, nomeadamente *S. carolitertii*, duas populações de *S. pyrenaicus* (Tejo e Guadiana), *S. torgalensis* e *S. aradensis*. Para este efeito combinámos novos dados de *RNA-seq* com dados de estudos anteriores para obter *assemblies* dos transcriptomas destas espécies e de *Leucisucs burdigalensis* e *Danio rerio*, usados como outgroups. Estes dados permitiram-nos caracterizar o número de genes ortólogos com assinaturas de seleção positiva*,* identificar os ramos da filogenia destas espécies com evidência de seleção positiva e determinar quais as funções biológicas com assinaturas de seleção ao longo da filogenia dos *Squalius* do oeste da Península Ibérica. Para além disso, dado que para detectar seleção é necessário inferir a uma árvore filogenética para cada gene ("*gene tree*"), foi possível caracterizar as relações filogenéticas entre estas espécies ao nível do transcriptoma.

Como usámos dados de *RNA-seq* de órgãos e condições distintas, foi necessário ter particular atenção à possibilidade de existirem diferentes isoformas nos alinhamentos de genes ortologos. Por este motivo,

desenvolvemos três *pipelines* bioinformáticas diferentes para 1) identificar grupos de sequências ortologas (i.e potenciais genes ortologos) entre os transcriptomas das nossas espécies; 2) criar alinhamentos de boa qualidade dessas sequências ortologas sem serem afetados pela presença de diferentes isoformas de *splicing*; e 3) realizar testes de seleção positiva nos alinhamentos limpos. Na primeira *pipeline*, Pipeline 1, utilizámos um script para implementar a metodologia do *Best Reciprocal Hits* (BRH) para a identificação de sequências ortologas e um método que agrupa sequências por similaridade para filtrar isoformas do mesmo transcrito. Na segunda *pipeline* que desenvolvemos, Pipeline 2, para a identificação dos genes ortologos utilizámos um pacote baseado no princípio do BRH mas com uma implementação mais completa (OrthoDB). Para a Pipeline 2 também desenvolvemos uma abordagem nova para lidar com alinhamentos com diferentes isoformas de *splicing* misturadas, que tem como princípio manter apenas as regiões do alinhamento onde as isoformas apresentam os mesmos exões. A última *pipeline* que desenvolvemos, Pipeline 3, é semelhante à Pipeline 2 mas é menos conservadora de modo a aumentar o número de genes ortologos identificados. Cada pipeline foi aplicada a um *dataset* de transcriptomas distinto. Os dados gerados pela Pipeline 3 foram os que utilizámos para as restantes análises. Dado que o o transcriptoma de *S. aradensis* era mais fragmentado do que o das outras espécies repetimos as análises com dois datasets que diferiam na inclusão (Dataset D) ou não (Dataset C) de *S. aradensis*.

Os nossos resultados demostraram que utilizando a Pipeline 1 fomos capazes de obter um número considerável de ortologos (10 767) e de ortologos com assinaturas de seleção (1 307). Contudo, a análise de alinhamentos mostrou que muitos dos genes sobre seleção positiva continham regiões desalinhadas, consistentes com a presença de exões de diferentes isoformas. Com a Pipeline 2 não encontrámos este problema, mas com o custo de uma grande redução no número total de ortologos (475) e de ortologos com assinaturas de seleção (19). Na Pipeline 3 obtivémos um número de ortologos comparável ao da Pipeline 1 (entre 9 605 e 13 525), e um menor número de ortologos com assinaturas de seleção (entre 106 e 247). Apesar destes resultados não serem diretamente comparáveis devido ao facto de não termos usado sempre os mesmo dados para as três *pipelines*, estes resultados sugerem que a Pipeline 3 consegue diminuir os falsos positivos devidos a *alternative splicing*, ao mesmo tempo que preserva o máximo de informação possível dos dados.

Relativamente aos padrões das *gene trees* nos transcriptomas das nossas espécies, descobrimos que no caso do Dataset C a *gene tree* mais suportada no transcriptoma (39% dos genes ortologos) corresponde à filogenia descrita recentemente com base em 7 genes nucleares, que sugere a parafilia de *S. pyrenaicus* em relação a *S. carolitertii*. Para além disso, uma outra porção do transcriptoma (10% dos ortologos) apoiava uma *gene tree* que segregava a espécies do clima Atlântico (*S. carolitertii* e *S. pyrenaicus* do Tejo) e do clima Mediterrâneo (*S. pyrenaicus* do Guadiana e *S. torgalensis*), o que pode ser um sinal de convergência nalguns genes entre espécies que habitam em regiões com o mesmo tipo de clima.

Relativamente às assinaturas de seleção, encontramos sinais de seleção em 2.0% dos genes ortologos no Dataset C e 1.4% no Dataset D. Isto não difere muito dos 2-4% de genes ortologos com assinaturas de seleção que estudos semelhantes estimaram noutras espécies de peixes ósseos. Em ambos os *datasets* encontramos sinais de seleção positiva em todos os pontos da filogenia, sendo que os ramos com mais genes com assinaturas de seleção positiva dentro do clade *Squalius* pertencem a espécies a viver sob a influência do clima Mediterrânico (*S. pyrenaicus* do Guadiana e ao *S. aradensis*). Visto que estudos prévios encontraram evidência de adaptação à temperatura noutra espécie que vive sob a influência do clima Mediterrânico, *S. torgalensis*, é possível que o relativo elevado número de genes sobre seleção positiva na linhagem destas espécies seja em parte uma resposta às altas temperaturas no Verão, características deste ambiente. No entanto, descobrimos que estas espécies de *Squalius* apresentam genes com assinaturas de seleção em funções tão variadas como a coagualação sanguínea, a resposta imunitária, desenvolvimento, proteólise, ligação entre proteínas e metabolismo. Nalguns casos, estas funções parecem estar sobretudo associadas a determinados ramos da filogenia. Por exemplo, a coagulação sanguínea tem 3 dos 5 genes desta categoria no ramo do *S. pyrenaicus* do Guadiana. Contudo, na maioria dos casos o número de genes sob seleção positiva em cada categoria funcional estavam distribuídos de forma idêntica ao longo dos ramos da filogenia. Isto indica que a nossa análise de enriquecimento funcional detetou funções biológicas que foram selecionadas de forma consistente ao longo da filogenia destas espécies. Em conclusão, os resultados do presente estudo sugerem que há 1.4-2.0% de genes sobre seleção positiva em todos os ramos da filogenia, principalmente em espécies sob a influência do clima Mediterrânico, que no geral apresentam um número maior de genes sobre seleção positiva.

Este estudo demonstra ainda que, em estudos comparativos que usem transcriptomas, a presença de *alternative splicing* pode facilmente afetar os alinhamentos, e eventualmente levar a falsos positivos. Nesta tese foi desenvolvida uma nova metodologia para lidar com a presença de isoformas de *splicing* alternativo nos alinhamentos, baseada no princípio de manter apenas os exões em comum entre isoformas de diferentes espécies. Esta pipeline bioinformática aqui desenvolvida é um recurso que pode ser útil para estudos comparativos em transcriptomas noutras espécies. Finalmente, relativamente aos *Squalius* do oeste da Península Ibérica, este estudo traz novas ferramentas transcriptomicas para duas destas espécies – *S. aradensis* e a população do Tejo de *S. pyrenaicus*, para as quais não existiam dados transcriptomicos. Os resultados deste estudo servem de ponto de partida para estudos mais detalhados no futuro.

Palavras-chave: Seleção positiva, dN/dS, transcriptomica comparativa, *splicing* alternativo, *Squalius.*

Abstract

One of the main goals of evolutionary biology is to understand the molecular mechanisms of adaptation. Advances on next generation sequencing (NGS) have allowed to improve our knowledge on the mechanisms of adaptation, including in non-model organisms. One example is the use of RNA-seq data to test at the transcriptome level for the presence of signatures of positive selection using the ratio of non-synonymous to synonymous mutations (dN/dS ratio). However, the identification of orthologous sequences between the transcriptomes of different species is challenging because of the possibility of mixing different splicing isoforms on the ortholog alignments. Even so, by providing tens of thousands of sequences for protein coding genes, *RNA-seq* can be a powerful tool for understanding the time and mode of the adaptative process.

In Portugal, the western Iberian freshwater cyprinids of the *Squalius* genus are a good system to study adaptation. The reason is that there are four species (*S. carolitertii, S. pyrenaicus, S. torgalensis* and *S. aradensis*) distributed across a north-south temperature cline, encompassing two distinct climate types – Atlantic and Mediterranean. Recent studies found evidences of adaptation to temperature in one of the southern species (*S. torgalensis*). In this study, we compared the transcriptomes of these four species to look for genes with signatures of positive selection, infer branches of their phylogeny with evidence for positive selection, and identify biological functions that were enriched in genes under positive selection. We also characterized the relationship between these species at the transcriptome level.

Since our *RNA-seq* data for the different species came from different organs our study was especially vulnerable to the effect of alternative splicing. We have thus developed a new approach to deal with alternative splicing in comparative studies using transcriptomic data. Our approach was based on identifying ortholog alignments with different splicing isoforms and remove the regions on the alignments with exons that were not common between isoforms. Our results suggest that our approach manages to reduce the quantity of false positives related to alternative splicing in comparison with a more conventional approach.

Regarding the phylogenetic relationship between species, we found support for the paraphyly between *S. pyrenaicus* and *S. carolitertii,* which has been also suggested by recent studies. Regarding the patterns of positive selection on these species, we found positive selection in 1.4% to 2.0% of the identified ortholog gene groups, which is comparable to what has been estimated for bony fish species in other studies. Interestingly, we found a relatively higher number of genes under positive selection on the branches of the southern species under the Mediterranean climate type than on the northern species under the Atlantic climate type. This could suggest that the southern *Squalius* species might be under stronger selective pressures due to the characteristics of the Mediterranean climate type, like high summer temperatures. We also found that the genes with signatures of selection were enriched on several biological functions,

including blood coagulation, immunity, proteolysis, development and metabolism. Rather than having particular functions associated with specific branches of the phylogeny, most of the biological functions were generic and distributed similarly across species. This suggests that these biological functions have been consistently selected on the phylogeny.

In conclusion, in this study we present a new approach to deal with alternative splicing on comparative studies using transcriptomic data, which can be useful for comparative studies on other species. We also present new transcriptomic data for two species of western Iberian *Squalius* – *S. aradensis* and the Tagus population of *S. pyrenaicus*. These results can be used as a resource for further studies on adaptation using the western Iberian *Squalius* as a model.

Keywords: Positive selection, dN/dS, comparative transcriptomics, alternative splicing, *Squalius.*

Table of Contents

List of Figures and Tables

Tables

Figures

1. Introduction

Adaptation is the evolutionary process through which phenotypic traits that increase the fitness of individuals on a given environment change due to the action of natural selection (Barrett and Hoekstra, 2011). This evolutionary mechanism has allowed living organism to colonize many of the different environment present on Earth. Therefore, it is not surprising that understanding the molecular mechanisms of adaptation is one of the main goals of modern evolutionary biology (Barrett and Hoekstra, 2011; Nielsen, 2005; Savolainen et al., 2013). Despite this, the molecular basis, the tempo and mode of adaptation are still poorly understood for most cases, especially for non-model organisms (Barrett and Hoekstra, 2011; Nielsen, 2005; Savolainen et al., 2013).

The revolution of next-generation sequencing (NGS) on the past years have contributed significantly to improve our knowledge on the mechanisms of adaptation by giving us access to genome-wide data (Barrett and Hoekstra, 2011; Savolainen et al., 2013; Stapley et al., 2010). This opens the door to understand the genetic basis of adaptation (how many and what genes?), and the genetic architecture of traits under selection (what regions of the genome?). Among other things, with next-generation sequencing we can study patterns at the genome level and hence we avoid the bias of choosing and focusing only on a few candidate genes (Barrett and Hoekstra, 2011), we have the power to identify adaptive mutations and to differentiate selective forces from demographic factors (Barrett and Hoekstra, 2011); and we are able to see how selection works not only at the level of genes, but also at the level of whole molecular networks (Daub et al., 2013; Foll et al., 2014). In recent years, the falling costs of sequencing have allowed to expand this type of studies to non-model organisms, namely ecological model organisms, for whom the main traits involved in adaptation are known but identifying their genetic basis was challenging due to a lack of genetic tools (Stapley et al., 2010).

Adaptation can be studied with a *bottom-up* approach, where we start by identifying putative adaptive genes through molecular signatures and then try to identify their phenotypic effects (Barrett and Hoekstra, 2011). Two common approaches to identify these putative adaptive genes are (i) population genetics and (ii) comparative genetics (Nielsen, 2005). Population genetics use data on the variation of the allelic frequency of polymorphic genetic markers within and between populations, to detect patterns consistent with the action of selection within or across populations. Comparative genetics, on the other side, uses data on the differences on gene sequence between species to detect patterns of past selection in one or more species. This last approach relies mostly, but not only, on comparing the number of non-synonymous and synonymous substitutions (dN/dS ratio) present in a gene to measure the strength and mode of selection acting on protein coding genes at a genomic scale (Jeffares et al., 2015). The idea behind this test, which is only applicable to the protein coding sequencing of a gene, is that non-synonymous substitutions (dN) are substitutions that change the aminoacid sequence of the protein and hence are considered potentially under

selection, while synonymous substitutions (dS) are substitutions that do not alter the aminoacid sequence of the protein and hence are considered neutral. When this ratio, also denominated ω, is less than one (ω<1), it indicates purifying selection, as the proportion of non-synonymous is lower than the synonymous substitutions; a ω = 1 indicates neutral evolution as both non-synonymous and synonymous substitutions occur at the same rate; and a ω >1 indicates positive selection as the rate of non-synonymous substitutions is larger than the synonymous substitutions (Jeffares et al., 2015; Nielsen, 2005). This statistic can be obtained at different levels from gene-wide to site-level, which can be used to answer different questions, such as: what genes are under purifying/positive selection? And in more detail, what are the sites under selection? Or, what are the branches in the gene tree under selection?

One type of NGS data for which the comparative genetics approach has been used to study adaptation is to sequence RNA of transcripts expressed at cells from different tissues, a technique known as RNA-seq that generates transcriptomic data (Baker et al., 2018; Cicconardi et al., 2017; Ghiselli et al., 2018; Yang et al., 2012; Zhao et al., 2014). Besides the great amount of gene transcripts present on the transcriptome, two other traits make this type of NGS data ideal for comparative analysis: 1) the fact that sequences from protein coding genes are easy to filter through the polyA tail of mRNA, and hence we can target gene sequences with high certainty; and 2) the fact the transcript sequences on the transcriptome are from processed mRNA and therefore, for each gene we obtain the final CDS sequence without intronic regions. Nonetheless using transcriptomic data for comparative analysis also presents a significant challenge for the identification of orthologous sequences due to the presence of alternative splicing (Breschi et al., 2017; Zambelli et al., 2010). Orthology identification is usually based on sequence similarity to identify ortholog sequences between species. Since splicing isoforms only diverge on the alternative spliced exons, different splicing isoforms of the same transcript can end up on the alignments of orthologous transcripts and bias the results of comparative analysis due to the presence of different exons between species. This is a very relevant issue considering that alternative splicing is a widespread phenomenon across the transcriptome, estimated to affect more than 90-95% of genes on humans (Pan et al., 2008; Wang et al., 2008; Zambelli et al., 2010). Some transcriptomics studies deal with alternative splicing by using methods to cluster very similar sequences on the transcriptome and then select only the longest or more expressed sequence of each cluster (Cicconardi et al., 2017; Jesus et al., 2016; Kong et al., 2014; Machado et al., 2015; Wang et al., 2017). However, this approach is based on the assumption that the expressed transcript isoforms are the same across all species. Considering the dynamic nature of the transcriptome, this might not be true for all of the transcripts, especially for studies using transcriptomes that come from public repositories where the organs, life stage, and conditions on which the transcriptomes were sequenced, and the protocols used for the sequencing of the transcriptomes are not necessary the same. Nevertheless, RNA-seq and transcriptomic studies allow to perform comparative genetic studies, even for non-model organisms with large genomes.

Two types of biological models are commonly used in studies of adaptation. One are species/populations on contrasting environments and with contrasting traits, and the other are species/populations distributed along environmental clines that exhibit a corresponding clinal variation on certain traits (Barrett and Hoekstra, 2011). Understanding the genetic basis of the differences between these populations/species can give us insights in to the molecular process of adaptation to their environments (Barrett and Hoekstra, 2011).

In Portugal, the cyprinids from the genus *Squalius* are a very well studied group of freshwater fishes that are distributed from North to South of the country along an environmental cline of temperature, which is reflected on a cline of external morphology traits, namely size, number lateral line scales, number of fin rays and number of gill rakers (Coelho et al., 1998). Currently, besides the *Squalius alburnoides* complex, there are four *Squalius* species in Portugal, *S. carolitertii*, *S. pyrenaicus, S.torgalensis* and *S.aradensis.* All of which are endangered except for *S. carolitertii* (Henriques et al., 2010; Machado et al., 2015; Mesquita et al., 2005). *S. carolitertii* is found in the Lima, Douro, Vouga and Mondego basins; *S. pyrenaicus* is found in the Samarra, Colares, Tagus, Sado, Guadiana, Almargem and Quarteira; *S. torgalensis* is found in the Mira drainage and *S. aradensis* in the Arade, Seixe and Quarteira drainages (Coelho et al., 1995, 1998), (Figure 1.1). Moreover, these species inhabit the two climatic types that exist in the Iberian Peninsula: Atlantic, with mild temperatures and a stable climate; and Mediterranean, which is characterized by strong seasonal cycles, with high temperatures and droughts in summer (Carvalho et al., 2010; Henriques et al., 2010; Magalhaes et al., 2003). Species at the north of the Iberian Peninsula like *S. carolitertii* and the Tagus population of *S. pyrenaicus* are under the influence of the Atlantic climate type, while species at the south like the Guadiana *S. pyrenaicus, S. torgalensis* and *S. aradensis* are under the influence of the Mediterranean climate type (Jesus et al., 2016, 2017).

Squalius species have been investigated in controlled conditions in the laboratory in recent studies (Jesus et al., 2016, 2017) to test the response of *S. carolitertii* and *S. torgalensis* to a predicted scenario of climate change for the end of the century, focusing on temperature and pH changes (Jesus et al., 2016, 2017). These two species inhabit the two climatic types, *S. carolitertii* in the Atlantic and *S. torgalensis* in the Mediterranean (Jesus et al., 2017). The authors studied a set of fourteen target genes, obtained from previous

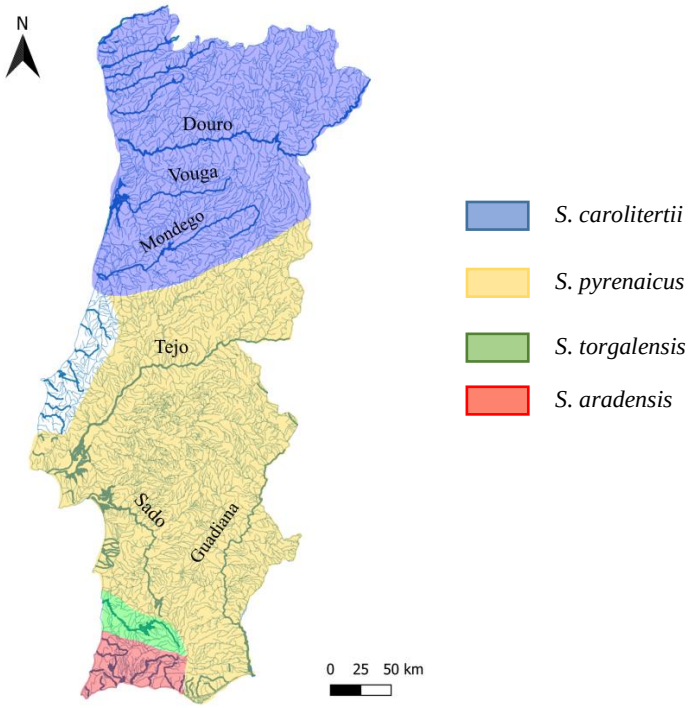


Figure 1.1 - Distribution of the four Squalius species in Portugal.

studies on acute thermal stress on these species (Jesus et al., 2016), using two different approaches: 1) measure gene expression response on individuals from both species exposed to different experimental controlled conditions in the laboratory of warming and/or acidification; 2) studying the predicted effects of the fixed substitutions on these genes between the two species in the structure and physicochemical properties of the proteins they code for. The authors found that the gene expression of these 14 genes was significantly different among temperature/pH conditions in *S. carolitertii* (the northern species) but that *S. torgalensis* (the southern species) gene expression was mostly unchanged. Together with these results they found that the predicted thermostability in proteins related with thermal response (HSP90) and with immunity (GBP1) is higher for *S. torgalensis*, and that there are structural differences in genes related with thermal response (HSC70 and FKBP52), immunity (GBP1) and glucose methabolism (HIF1α). These results led the authors not only conclude that *S. torgalensis* has a greater tolerance to warming and acidification than *S. carolitertii*, but also to postulate that such increased tolerance result from adaptation to higher temperatures characteristic of Mediterranean climate, resulting from changes at the protein level on *S. torgalensis.* Even though the authors did not perform any type of selection tests based on molecular data (namely dN/dS) to test this hypothesis, other data from the group has shown evidences of adaptive convergence for some genes in the southern species (Coelho et al., unpublished data).

Taken together, the evidences of molecular adaptation found in previous works, the distribution of these species along an environmental cline, their clinal variation on traits of external morphology and the existence of some genomic tools, make this group of freshwater fishes an interesting model for the study of molecular adaptation. Therefore, the main goal of this thesis is to study the signatures of positive selection across the transcriptome and the phylogeny of Western Iberian *Squalius* freshwater fishes. To accomplish this, we defined three specific goals:

1) To develop a bioinformatics pipeline to identify orthologous genes across the transcriptomes of different species and use them to test for positive selection. The aim is that this pipeline takes into account alternative splicing and duplicated genes, thus enabling us to use transcriptomic data from different species obtained from different organs for comparative sequence analysis.
2) To characterize, for the first time, the signatures of positive selection across the phylogeny of these species at the transcriptome-level, using a branch-site dN/dS test to look for signatures of positive selection not only at the tips of the species tree (i.e. on the branches of extant species), but also on their ancestor species (i.e. the internal branches of the species phylogeny). For these tests we also inferred gene trees for each orthologous gene, allowing us to characterize the transcriptome-wide relationships between species.
3) To use functional enrichment analysis to gain insights about the molecular and biological functions that were under positive selection through time across the phylogeny of these species. We tested if biological processes of genes under positive selection were associated with particular branches of the species tree, to obtain a general view about adaptation of the *Squalius* clade through time.

This comparative transcriptomic study allowed us to identify the genes and biological functions that were involved in adaptations at different points in the evolutionary history of these species. Besides its fundamental relevance, understanding the action of natural selection through the phylogeny can also have practical implications, like helping to predict their potential response to future environmental changes.

2. Methods

2.1 RNA Seq data

In this study we combined newly generated RNA-seq data with available RNA-seq data from previous studies (Table 2.1). Regarding the new data, it includes data from the Iberian freshwater fish *Squalius aradensis* and *Squalius pyrenaicus* from Tagus and Guadiana populations. Regarding the available data from previous studies, these included libraries of: *S. carolitertii* and *S. torgalensis* (Genomic Resources Development Consortium et al., 2015a); *S. pyrenaicus* from the Guadiana population (brain and gonads) (Genomic Resources Development Consortium et al., 2015b); *Leuciscus burdigalensis* (Genomic Resources Development Consortium et al., 2015c), which was used as a close outgroup since *Leuciscus* is a sister clade to *Squalius* (Stout et al., 2016); *S. alburnoides* from the AA genomotype (Matos et al, in prep), whose genome should be similar to the genome of its *Anaescypris*-like parental, and hence was used as another outgroup; and finally the *Danio rerio* CDS (Howe et al., 2013) ,which was used as the basal outgroup, as well as the source of gene annotations. It is noteworthy that for simplicity, we will be using the term "species" for all of these samples including the Tagus and Guadiana populations from *S. pyrenaicus*. This is also because recent studies suggest that these two populations are likely paraphyletic in relation to *S. carolitertii* (Sousa-Santos et al., 2019; Waap et al., 2011).

2.1.1 Newly sequenced transcriptomes

2.1.1.1 - Sampling

The *S. aradensis* specimens were captured by electrofishing (performed with low duration pulses to avoid killing juveniles, 300V and 2-4A) from the Odelouca stream in the Arade basin (37º17'0.53"N; 8º29'7.31"W). The individuals were anesthetised and euthanized with an overdose of tricaine mesylate (400 ppm of MS-222; Sigma-Aldritch, St. Louis, MO, USA). Organs were stored in RNAlater® at -80ºC until further use. We also used tissue samples from individuals of *S. pyrenaicus* previously sampled on the Tagus and Guadiana basins (Matos et al., 2015) which were preserved at -80ºC. The fishes from the Tagus population came from the Portuguese river Ocreza from the Tagus basin (39º43'48.23"N; 7º45'38.13"W), while the fishes from the Guadiana population came from de Oeiras stream from the Guadiana basin (37º37'30.30"N; 7º48'37.03"W) (Matos et al., 2015).

2.1.1.2 - RNA extraction

RNA was extracted from four individuals for each of the three samples from *S. aradensis* and *S. pyrenaicus* mentioned above. For both *S. pyrenaicus* populations RNA was extracted from muscle. For *S. aradensis*, RNA was extracted from brain tissue because it was the only tissue available in good conditions for this species. Tissue Ruptor (Qiagen, Valencia, CA, USA) was used to homogenize the organs and afterwards RNA was extracted using the Total RNA Purification Kit (Norgen Biotek Corp., Thorold, ON, Canada) according to the manufacturer's instructions. RNA concentration was measured on Qubit® 2.0 Fluorometer (Thermofisher Scientific, Waltham, MA, USA) and its quality (RNA quality number – RQN and 28S/18S) evaluated using AATI Fragment Analyzer (Advanced Analytical Technologies, Inc.). Once confirmed the good quality of the RNA libraries, these were pooled per species and then stored on dry powder at room temperature using the GenTegra-RNA Kit (GenTegra® LLC., Pleasenton, CA, USA) for shipment to the Beijing Genomics Institute (BGI, Hong Kong, China) to be sequenced. Once at BGI, Agilent 2100 Bioanalyzer (Agilent RNA 6000 Nano Kit) was used to confirm the quality of the samples (measuring the RNA concentration, RIN value, 28S/18S and fragment length distribution).

2.1.1.3 - Library preparation, sequencing and read cleaning

Library preparation and sequencing was performed in outsourcing at BGI as follows: 1) mRNA was isolated from total RNA using the oligo(dT) method to isolate poly(A) RNA; 2) mRNA was fragmented; 3) first and second cDNA strands were synthesized; 4) cDNA fragments were purified and resolved with EB buffer for end reparation and poly(A) tail addition; 5) cDNA fragments were connected with adapters; 6) cDNA fragments of suitable size were selected for PCR amplification; 7) the quality and quantity of those libraries was evaluated using Agilent 2100 Bioanalizer and ABI StepOnePlus Real-Time PCR System; and 8) finally libraries were sequenced using Ilumina HiSeq 4000 to generate 100 bp paired-end reads.

The sequencing resulted in 11.5 GB of raw reads, which were filtered at BGI by discarding reads with adaptor sequences, reads with more than 5% of unknown bases (N) and reads with 30% or more of the bases with a Phred base score lesser than 15. This resulted in 11 GB of cleaned reads that together with the raw reads were stored in FASTQ format. Once we received this data, we confirmed the quality of the clean reads independently using FastQCv0.11.7 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

Table 2.1 – Transcriptomic data for each transcriptome assembly used in this study.

Species	Read Library Source:	Tissues	Number of Individuals	Pooled by tissue	Dataset	Type of assembly	Assembly label
S. carolitertii	Genomic Resources Development Consortium et al., 2015b	muscle, liver and fins	7	yes	A and B	Original published assembly	ScarolO
					C and D	re-assembled	ScarolR
S. pyrenaicus (Tagus)	This study	muscle	4	yes	C	*De novo* assembly (preliminary version)	SpyrenTP
					D	*De novo* assembly (final version)	SpyrenTF
S. pyrenaicus (Guadiana)	Genomic Resources Development Consortium et al., 2015c	Brain and testis/ovary	1 male and 1 female	no	B	Original published assembly	SpyrenGO
					A and C	Re-assembled	SpyrenGR
	This study / Genomic Resources Development Consortium et al., 2015c	muscle / brain	4 / 2	yes / no	D	de *novo* assembly with mixed libraries	SpyrenGD
S. torgalensis	Genomic Resources Development Consortium et al., 2015b	muscle, liver and fins	7	yes	A and B	Original published assembly	StorgalO
					C and D	re-assembled	StorgalR
S. aradensis	This study	brain	4	yes	D	*de novo* assembly	Sarad
S. alburnoides	(Matos et al, in prep)	juvenile	4	yes	A	*de novo* assembly	Salbur
L. burdigalensis	Genomic Resources Development Consortium et al., 2015a	spleen, head kidney, fins	4 healthy and 4 infected	no	A and B	Original published assembly	LburdiO
		spleen, head kidney, fins	2 healthy	no	C and D	Re-assembled with libraries of only two individuals	LburdiR

2.1.2 – Published transcriptomes

We had direct access to cleaned reads for *S. carolitertii*, *S. torgalensis* and *S.pyrenaicus* from the Almargem River (Guadiana basin), available from previous studies (Genomic Resources Development Consortium et al., 2015a, 2015b) on the following online repositories: *S. carolitertii* and *S. torgalensis* on NCBI SRA, under the project accession numbers SRP049802 and SRP049801, respectively; and *S. pyrenaicus* on the European Nucleotide Archive under the study accession number PRJEB9465. We also had access to cleaned reads from *S. alburnoides* AA form from an unpublished RNA-seq study (Matos et al, in prep). Cleaned reads from healthy *L. burdigalensis* individuals were downloaded from NCBI SRA with the project accession number SRP049407 (Genomic Resources Development Consortium et al., 2015c), using "fastq-dump" from the SRA-Toolkit v2.8.2.1 (Leinonen et al., 2011). Finally, we also downloaded the the *Danio rerio* database of predicted coding sequences by Ensembl, using the GRCz10 genome assembly of *Danio rerio* (Howe et al., 2013) .

2.2 – Datasets

Four different transcriptome datasets were used during this study. The different datasets correspond to the chronological order in which they were analysed due to the timing at which we received the RNA-seq data from BGI (the latest being received on September 2018). Moreover, based on results we were getting we also made some adjustments in the datasets through time. The four datasets used were (Table 2.2):

- **Dataset A**, where we prioritized transcriptome assemblies created with the same software (Trinity). When species had no published transcriptome assembly or, if available, were not assembled with Trinity, we performed a *de novo* assembly. We used the following data: 1) three published transcriptome assemblies from *S. carolitertii* (ScarolO), *S. torgalensis* (StorgalO) and *L. burdigalensis* (LburdiO); 2) one *de novo* re-assembly of the published transcriptome of the Guadiana *S. pyrenaicus* (SpyrenGR); 3) One *de novo assembly* of the unpublished *S. alburnoides* from the AA genomotype (Salbur); and used 4) the *D. rerio* CDS database as outgroup and source of gene annotations. For dataset A we ensured that all datasets were generated with the same software, using default settings.
- **In Dataset B,** we changed the outgroup species and changed the approach by using the transcriptome assemblies as published in their original studies. Thus, Dataset B was the same as Dataset A but without the *S. alburnoides* and using the published Guadiana *S. pyrenaicus* assembly (SpyrenGO), instead of our Trinity re-assembly.
- **In Dataset C,** we included new data and *de novo* re-assembled all the transcriptomes ourselves using Trinity with exactly the same settings. We 1) re-assembled the transcriptomes of *S. carolitertii*

(ScarolR), *S. pyrenaicus* from Guadiana (ScarolGR), *S. torgalensis* (StorgalR) and *L. burdigalensis* (LburdiR). In the case of *L. burdigalensis* we only assembled libraries from 2 out of 8 individuals due to computational limitations; 2) *de novo* assembled the preliminary RNA-seq data of the newly sequence Tagus *S. pyrenaicus* transcriptome; 3) we included the *D. rerio* coding sequence (CDS) database to use as outgroup and source of gene annotations.

- **Dataset D** was similar to C, but included all the newly sequenced RNA-seq data. It differed from Dataset C on: 1) the Tagus *S. pyrenaicus*, since we used the final RNA-seq data (SpyrenTF) instead of the preliminary data; 2) the Guadiana *S. pyrenaicus*, since we merged the published brain library with our newly sequenced muscle library (SpyrenGR); 3) the new RNA-seq data of *S. aradensis.* Just as in Dataset C, all the published transcriptomes were *de novo* re-assembled using exactly the same settings as used for the *de novo* assemblies.

Table 2.2 - Summary of the species and transcriptome assemblies used on each dataset. The symbol "-"means that species was not used for that dataset.

Species	Dataset A	Dataset B	Dataset C	Dataset D
S. carolitertii	ScarolO	ScarolO	ScarolR	ScarolR
S. pyrenaicus (Tagus)	-	-	SpyrenTP	SpyrenTF
S. pyrenaicus (Guadiana)	SpyrenGR	SpyrenGO	SpyrenGR	SpyrenGD
S. torgalensis	StorgalO	StorgalO	StorgalR	StorgalR
S. aradensis	-	-	-	Sarad
S. alburnoides (AA)	Salbur	-	-	-
L. burdigalensis	LburdiO	LburdiO	LburdiR	LburdiR

2.2 – Bioinformatic pipelines for identification of orthologs with signatures of selection

To detect signatures of positive selection on orthologous sequences without being affected by high rates of false positives or low number of orthologs, we needed a bioinformatic pipeline to 1) identify of orthologous sequences between the transcriptomes of our species, 2) obtain good quality clean alignments of the orthologous sequences, and 3) perform comparative analysis on these alignments, namely positive selection tests. We developed three different pipelines to achieve this, each one improving on the previous one. For simplicity we named these pipelines per order of development as Pipeline 1, Pipeline 2 and Pipeline 3 (Figure 2.1). Pipeline 1 used simple approaches for ortholog identification and transcript isoform filtering that are commonly used on comparative studies, and was loosely based on the bioinformatic pipeline used by Wang et al., to do phylogenetic analysis on the transcriptome of four luminescent beetles (Wang et al., 2017). Pipeline 2 used a more sophisticated method for ortholog identification than Pipeline 1 and dealt better with splicing isoforms, but it was too conservative. Pipeline 3 was based on Pipeline 2 but was less conservative while still dealing with the splicing isoforms, having several improvements and corrections compared to Pipeline 2. Pipeline 3 was the final version of our bioinformatic pipeline, and the dataset of orthologous sequences with signatures of selection identified with this pipeline was used for further analysis (see below). For the chronological reasons mentioned previously, for each pipeline we used different transcriptome datasets: for Pipeline 1 we used Dataset A; for Pipeline 2 we used Dataset B; and for Pipeline 3 we used Dataset C and D. Next, we describe the various steps that composed these pipelines and how they differ from each other, giving special emphasis to Pipeline 3.

2.2.1 – *De novo* transcriptome assembly

We performed *de novo* assembly of transcriptomes in Pipeline 1 (Dataset A) and in Pipeline 3 (Dataset C and D), i.e. we assembled the transcriptomes without using a reference genome. We used Trinity v2.5.1 (Grabherr et al., 2011) with default settings (kmer size = 25, minimum kmer coverage = 1, minimum contig length = 200 bp). Regarding the process of assembly, Trinity algorithm involves three steps. First, find the full sequence of the dominant isoform of a transcript and the unique sequences of the alternative isoforms. Second, cluster those sequences and construct a de Brujin graph per gene, or cluster of similar genes. Finally, process these graphs to find overlapping sequences to infer the full sequence encompassing all isoforms of a transcript, attempting to distinguish isoforms from paralogous genes. To avoid any potential bias of comparing transcriptomes assembled with different versions of Trinity, different settings or even different assemblers, on Pipeline 3 (Dataset C and D) we also *de novo* re-assembled all the species that had published

transcriptome assemblies using the same settings as above. Because our goal was to obtain homologous gene sequences across species ignoring gene expression levels, pooling data from different organs did not pose any problems to our analyses. Thus, to increase the power of the assembly algorithm, we assembled together all the RNA-seq libraries from each species, independently of whether they came from different tissues or samples, following the suggestion on Trinity's manual. It is noteworthy that to improve the coherence of the organs used on our assemblies, on Dataset D we re-assembled the *S. pyrenaicus* from Guadiana without the gonads library and adding our newly sequenced muscle library.

In all the three pipelines, we included a step to evaluate the completeness and quality of the resulting transcriptome assemblies. This was done using BUSCO v3.0.2b (Simão et al., 2015; Waterhouse et al., 2017) and TransRate v1.0.3. (Smith-Unna et al., 2016). BUSCO assesses the completeness of a genome or transcriptome assembly by comparing it against databases of universal single-copy orthologs for a given evolutionary lineage. For the assessment of the completeness of our six transcriptomes we selected the Actinopterygii (ray-finned fishes) database, which was the most specific lineage available in BUSCO that encompassed our species. TransRate is a tool used to evaluate the quality of *de novo* transcriptome assemblies using only the reads and the assembly as input, and computing several quality metrics, such as N50, which can be used as a quality metric that gives an estimate of the distribution length of the resulting assembled fragments (contigs).

2.2.2 – Contig redundancy removal (Unigene identification)

A common step in transcriptomic studies data is the obtention of a unigenes dataset, i.e a dataset with only one sequence representing each gene per species. The goal is to filter all but one isoform per transcript (usually the longest or the most expresed) in order to simplify downstream analysis. For Pipeline 1 and Pipeline 2 we used CD-HIT-EST v4.6.8 (Fu et al., 2012), which implements a common approach of clustering sequences by similarity, keeping only the longest contig for each cluster. We ran CD-HIT-EST with a similarity threshold value of 0.80 for Pipeline 2, since in our testing this value seemed to eliminate most of the alternative splicing on the transcriptome assemblies. However, since CD-HIT-EST takes longer to run with lower values of similarity threshold, on Pipeline 1, due to time constrains we were only able to run CD-HIT-EST with a similarity threshold of 0.80 for the Guadiana *S. pyrenaicus* assembly, and 0.85 on the *S. torgalensis.* On Pipeline 3 we used a more sophisticated step to jointly detect unigenes and orthologous during the ortholog identification step (see below), and hence this step was not required.

2.2.3 – Open Reading Frame (ORF) identification

We built a CDS database for all the transcriptome assemblies in a given dataset by predicting the Open Reading Frames (ORFs) of all the unigenes on the assemblies. This was done using TransDecoder v5.0.2 (https://github.com/TransDecoder/TransDecoder/wiki), considering a minimum sequence size of 300 bp and running the TransDecoder.LongOrfs and TransDecoder.Predict modules. In order to be able to use this CDS dataset for downstream analysis we trimmed the TransDecoder header and simplified the sequences headers using a custom Bash script (Supplementary Folder 2.1).

2.2.4 - Ortholog identification and annotation

To identify orthologs groups (OG) of sequences between the transcriptomes and *Danio rerio* CDS databases we used two different approaches. For Pipeline 1 we wrote custom Bash and R scripts (Supplementary Folder 2.1) that implemented a Best Reciprocal Hits (BRH) methodology to find OGs using tools from BLAST 2.6.0 + (Camacho et al., 2009). This approach is based on the simple principle that orthologous sequences across two transcriptomes should be reciprocal best matches of each other when using local alignment search tools like BLAST. It also assumes that if there are other sequences with comparable similarity scores, then they are very likely paralogous, and hence those sequences should be removed from further analysis. However, BRH does not guarantee that paralogous genes are discarded if for example the true ortholog of a sequence in one transcriptome was not sequenced on all the other transcriptomes. In that case paralogous genes from those transcriptomes could be the best reciprocal hits.

In order to increase our confidence that the dataset contained only true orthologous, reducing the possibility of including paralogous, we considered a more sophisticated approach. Thus, for Pipeline 2 and 3 we used the OrthoDB suite v2.3.1, which is based on the pipeline used for the creation of the database of ortholog groups ORTHODB v.9.1. (Zdobnov et al., 2017). This pipeline works by: 1) finding the best reciprocal hits (BRH) of genes between assemblies; 2) searching for BRH within the assemblies that are more similar than the BRH between assemblies – these are called in-paralogs and should represent gene duplications after the speciation event; 3) combine the results of the two previous steps to create clusters of orthologous sequences, the orthologous groups (OGs), which theoretically should descend from a single gene in the ancestral of all the species (Kriventseva et al., 2015). We ran the OrthoDB pipeline with default settings except that we used Blast+ v2.7.1 (Camacho et al., 2009), which is a faster alignment algorithm than the default.

We processed the output of ORTHODB with custom scripts, which allowed us to jointly detect orthologous unigenes. First, for each ortholog group identified, we obtained the OrthoDB IDs of all the

sequences belonging to that group. Then, for each species, those OrthoDB IDs were mapped to the corresponding sequence ID on the transcriptome assembly. We only considered ortholog groups with at least one sequence per species. Due to isoforms, some orthologs groups had more than one sequence per species, and in these cases we kept only the longest sequence per species. Because we were working with the transcriptome instead of the genome, these similar sequences were most likely isoforms rather than gene duplications. By choosing the longest sequence we aimed to keep the most informative isoforms for the downstream analysis to detect selection. It is important to notice that despite of this approach being more sophisticated than a simple BRH, and in theory being more accurate, there is still no guarantee that we kept the same isoform for each species in every OG identified. We addressed this potential confounding factor in a trimming step after the alignment of the OG sequences from different species (see below).

Finally, we created a BLAST custom database for each transcriptome with the *makeblastdb* module from Blast+ v2.7.1. Given the transcriptome ID of each ortholog group we obtained the sequences from the transcriptomes databases with the *blastdbcmd* module of Blast+, as a fasta file per ortholog group with the unaligned sequence of each species. For each ortholog group we assumed that the *Danio rerio* sequence annotation corresponded to the annotation of sequences on the ortholog group.

2.2.5 – Ortholog group (OG) sequence alignment

Having our dataset of orthologous groups (OGs) between our transcriptomes, we proceeded to aligning them. In Pipeline 1 and Pipeline 2 we aligned the sequences of the OG using ClustalOmega v.1.2.4 (Sievers et al., 2011) with default settings. In Pipeline 3 we performed this step using T-COFFEE v11.00 (Notredame et al., 2000). This is because T-COFFEE preserved the integrity of the ORFs better than ClustalOmega, since it performs the nucleotide alignment accounting for the protein sequences (options -other_pg seq_reformat -action +translate command on T-COFFEE). We aligned our sequences using the fM-COFEE algorithm, a fast version of the M-COFFEE algorithm (Wallace, 2006) which only uses three fast aligners to create a consensus: MUSCLE (Edgar, 2004), MAFFT (Katoh et al., 2002) and Kalign (Lassmann and Sonnhammer, 2005). Finally, we used the "-other_pg seq reformat -action +thread_dna_on_prot_aln" options to map the nucleotide sequences on the protein alignments and obtain the multispecies nucleotide alignments for each OG, with one sequence per species.

2.2.6 – Cleaning orthologous group (OG) alignments

After obtaining the OG alignments, we used used GBLOCKS v0.91b (Castresana, 2000; Talavera and Castresana, 2007) to remove gaps and poor quality alignment regions. GBLOCKS algorithm involves the following steps: 1) finding blocks within the alignment with conserved flanking regions; 2) removing regions with too many contiguous nonconserved positions; 3) by default, removing gap positions; 4) removing blocks that are too small. For Pipeline 1 and Pipeline 2 we ran GBLOCKS in codon mode, with all settings on default, namely not allowing gap positions on the blocks, a minimum block length of 10 bp, and a maximum number of contiguous nonconserved positions of 8 bp. For Pipeline 3 we ran GBLOCKS with the same settings, except for the minimum block length which we changed to 50 bp because we found that a lower value allowed short conserved blocks within poor quality regions to be kept on the aligments, which could lower the overall quality of the alignment.

GBLOCKS algorithm works by detecting the overall conservation of regions of the alignment. This has the undesired effect that if few individual sequences are completely misaligned, that region could still be considered conserved by the algorithm provided that the other sequences are conserved (Castresana, 2000). This can lead to incorrectly keeping misaligned OGs, which in downstream analyses could result in false signatures of positive selection. A careful visual inspection of random alignments indicated that this happened in some of our alignments. The misalignments however did not occur randomly across the whole alignment, but were mostly on certain regions, usually at the ends of the alignments, where one or two sequences suddenly diverged completely from the other species. Assuming this could be the result of a low quality of some sequences at the ends, we tried to remove such regions by creating a custom script to trim the ends of the alignments, removing 21 bp at each end (Pipeline 1). However, we found that the misaligned regions could have far more than 21 bp of length and that misalignments could also occur in the middle of the sequence. Notoriously, there were cases when two sequences were misaligned on the same region but were aligned between themselves. These patterns resemble what we would expected if the alignments comprised different splicing isoforms, suggesting that at least some of our OGs comprised different splicing isoforms across species. Considering that our transcriptomes came from different organs for different species, this is not surprising. Technical problems like incomplete sequencing of all the transcripts isoforms could also cause this situation. Since the algorithms of both our BRH script and OrthoDB are based on sequence similarity, in transcripts where a common isoform for all species did not exist, the more similar isoforms from each species are selected for each orthologous group, even if they were not from the same isoforms. This would result in OG alignments that would be misaligned on the regions were the alternative exons of different isoforms meet. In cases where this happened only for a couple of sequences, GBLOCKS could still fail to filter such misaligned regions. Due to the heterogeneity of the tissues where our

transcriptomes came from, we suspected that most OGs could contain a mixture of different splicing isoforms. For this reason, we decided not to remove all the OGs with misalignments because such an approach would potentially remove most of our OG dataset. Therefore, we developed an alternative approach as described next.

2.2.6.1 – A new approach to align OGs accounting for different isoforms across species

To remove poorly aligned regions due to a mixture of different splicing isoforms we created a custom R script. This complements GBLOCKS, by removing poorly aligned regions from the alignment that GBLOCKS was unable to detect. Assuming that these misalignments correspond to exons only found in some species (i.e. a different isoform), by discarding such regions we would thus obtain a multispecies alignment comprising only the common exons across species in the orthologous group.

The custom R script to minimize the impact of isoforms that we created involves the following steps. First, we calculate a matrix of pairwise differences (using the ratio of mismatches) between the sequences of all species against each other. These distances were calculated on windows of 48 bp along the alignments. Second, we computed the distribution of the pairwise mismatch distances between a given species and all others, across all the orthologous group alignments. Third, to detect misaligned regions, we considered that windows of 48bp with distances larger than a given quantile of the empirical distribution were outliers (0.95 for Pipeline 2 and 0.99 for Pipeline 3). On Pipeline 2, we removed regions with 3 consecutive windows of 48 bp whose pairwise mismatch distance value was superior to the 0.95 quantile for at least one species. On Pipeline 3, we took a more stringent filter removing all windows of 48 bp whose pairwise mismatch distance was larger than the 0.99 quantile at least for one species, regardless of their location on the alignment. Finally, we trimmed the first and last 18 bp of the remaining alignment to eliminate any tips of alternative exons that could still remain on the alignment. To keep only sequences with a reasonable number of sites, all resulting cleaned alignments with less than 100bp were discarded, which was done with a custom script. This is a conservative approach aiming to avoid any false positive signature of positive selection in downstream analyses due to alignment errors. However, by minimizing the probability of false positives we also increased the chances of discarding true positives, i.e. the risk of discarding divergent windows due to positive selection. The algorithm implemented in our custom R script was validated by analyzing simulated datasets where we introduced alignment errors at specific regions affecting mainly one species.

2.2.7 Inferring gene trees for aligned multispecies orthologous groups

Once we obtained cleaned alignments, we inferred the gene tree of each OG. For Pipeline 1 and Pipeline 2, gene trees for each OG clean alignment were estimated using the Neightbour-Joining module from HyPhy v2.3 (Pond et al., 2005) software package. Distances between sequences on the cleaned alignment were estimated using the Tamura-Nei (93) distance (Tamura and Nei, 1993), without allowing for negative branches lengths but allowing for unequal character frequencies, transversional bias corrections, and a gamma distributed (alpha = 1) rate variation from site-site. For Pipeline 3 we used the method implemented in FastTree v2.1.10 (Price et al., 2010), as this is based on approximately-maximum-likelihood approach. We used the default settings to calculate for each cleaned OG alignment a gene tree. Distances between sequences were estimated using the Jukes-Cantor distance matrix (Jukes and Cantor, 1969). FastTree works by first creating an initial neightbour-joining tree and then refining it heuristically with 11 rounds of minimum-evolution nearest neighbor interchanges (NNI), 2 rounds of subtree-prune-regraft (SPR) moves (also minimum evolution) and 11 rounds of maximum-likelihood NNIs.

2.2.8 - Detecting positive selection

For all the cleaned alignment of OGs we tested for molecular signatures of positive selection. OGs with evidence of positive selection are hereafter referred to as pOGs ("positive ortholog groups"). In particular, we aimed to characterize the timing of positive selection along the phylogeny of Iberian *Squalius* freshwater fish. The aim was to test if there was a particular branch of the species tree with more genes under positive selection (time), and to test if different genes and pathways were selected during a particular branch of the phylogeny. To detect signatures of positive selection we used aBSREL v2.0 (Smith et al., 2015), implemented on HyPhy, which is a branch-site method based on the ratio of non-synonymous and synonymous mutations (dN/dS ratio). In order to identify branches in the gene trees with evidence of positive selection, we used the aBSREL (adaptive Branch-Site Random Effects Likelihood) method, which is based on a branch-site model that tests for positive selection on each sequence separately and on the reconstructed ancestral sequences. The aBSREL models account for both branch-level and site-level heterogeneity on the dN/dS ratio (also known as ω), and allows testing for each branch whether a proportion of sites have evolved under positive selection ($\omega > 1$). The method aBSREL is different from other "branch-site" models in that it uses AIC_C (small sample AIC – Akaike Information Criterion) to infer the optimal number of ω classes for each branch, instead of using a fixed number of ω classes like in other models. Thus, aBSREL takes into consideration the possibility of varying evolutionary complexity between branches. To test for positive selection in a given branch, aBSREL fits a "full adaptive" model were $\omega>1$

classes are allowed, and through a Likelihood Ratio Test, compares it with a null model without $\omega > 1$ classes for each branch of the phylogeny. For each cleaned OG alignment we gave as input to aBSREL its respective alignment and inferred gene tree (see above). Since the aBSREL model does not account for stop codons, all stop codons were removed from the alignments with a custom script. We used custom scripts to process and summarize the individual reports of aBSREL for every OG and considered them to be under positive selection (pOGs) if the aBSREL test was significant for at least one branch in the corresponding gene tree (p-value < 0.05).

2.3 Downstream analyses

Once we obtained the final set of orthologous groups (OGs) and positive selection OGs (pOGs) with Pipeline 3 for Dataset C and D, we did further downstream analyses to accomplish the other two goals of the study: characterize the signatures of natural selection across the phylogeny and get insights on the biological functions that were selected through this clade (Figure 2.2).

2.3.1 Transcriptome-wide distribution of gene trees

We analyzed the distribution of gene tree across OGs to evaluate the variation of the relationship between species across the transcriptome. We expected most gene trees to be consistent with the most recent inferred species tree phylogeny for these species (Sousa-Santos et al., 2019; Waap et al., 2011). However, incongruences between the gene tree and species tree are also expected and can be due to neutral processes, such as incomplete lineage sorting (ancestral polymorphism) and/or introgression. These are likely to occur when working with closely related species, such as the four *Squalius* species (Waap et al., 2011) . The action of natural selection can also affect gene tree topology and lead to incongruences between the gene tree and species tree topology (e.g. due to convergence). Therefore, we quantified the proportion of OGs with gene trees compatible and incompatible with the inferred phylogeny for our species.

To cluster gene trees with similar topologies and characterize their distribution across the transcriptome we used the R package *treespace* (Jombart et al., 2017). This package uses tree metrics and multivariate analysis to: 1) perform a low-dimensional representation of gene tree variablility using principal coordinate analysis (PCoA); 2) identify clusters of similar gene trees; 3) create cluster-specific consensus gene trees. We used the Robinson-Foulds metric to compute distances between trees and the default Ward method to create clusters of gene trees. The number of clusters depended on the number of ingroup species on the dataset. As the number of possible topologies increases with the number of species, for Dataset C we

considered 5 clusters (4 ingroup *Squalius* species) and for Dataset D we considered 15 clusters (5 ingroup *Squalius* species). The gene tree representative of each cluster was obtained with the function "medTree" that calculates the geometric median tree of each cluster. For all pipelines we estimated the number of OG gene trees assigned to each cluster to quantify variation in gene trees across the transcriptome. For each dataset, we used the median tree of the most supported gene tree cluster as our reference species tree for subsequent analyses. We used our inferred species tree instead of the species tree recently reported because phylogenetic studies of the western Iberian *Squalius* disagree on some aspects, namely on a possible paraphyly between *S. carolitertii* and *S. pyrenaicus*. Therefore, to avoid biasing our results by choosing one of the species trees reported in previous studies, we decided to use our inferred species tree for subsequent analysis. We considered the species tree to correspond to the most frequent gene topology across gene trees. In Dataset C our inferred species tree showed paraphyly between *S. carolitertii* and *S. pyrenaicus*, while in Dataset D it showed a polytomy of *S. carolitertii* and *S. pyrenaicus* from Tagus and Guadiana (*see* Results).

2.3.3 – Mapping the signatures of positive selection in to the phylogney

To characterize the biological functions targeted by natural selection across the phylogeny of our species, we mapped the gene tree branches under selection on the pOGs into our inferred species tree. As expected, many gene trees were incongruent with the species tree. Hence, in those cases there was no direct match between the gene tree branches and the species tree. In those cases, to map a given gene tree inner branch into the species tree we performed two steps. First, we identified all the external branches downstream of that inner branch (i.e. all species downstream of that branch). Then, we found the most recent ancestor of those species in the species tree. For instance, if a pOG showed signatures of selection on the branch representing the immediate ancestor of *S. aradensis* and *S. carolitertii*, we would map this selective event into the most recent common ancestor between *S. aradensis* and *S. carolitertii* in the species tree, which in this example would be the ancestral of all *Squalius* species. Finally, using a custom R script we did a regression analysis to test if there was any correlation between the strength of the positive selection signal (p-value) and the length of the external branch.

2.3.2 - Functional enrichment analysis

For Pipeline 3 we included a step to perform a functional analysis aiming to test if the pOGs we found tended to be involved in particular biological functions. For each pOG, we assumed that the gene annotation of the *Danio rerio* was valid for all other species. We tested for significant enrichment of the functional annotations of these pOGs genes. We used annotations from many different databases, namely Gene

Ontology, UniProt, InterPro, OMIM, Reactome, KEGG Pathway, SMART, Pfam, PROSITE, COG, BioCarta and SwissProt. We compared pOGs with evidence of positive selection in different branches of the gene trees, aiming to infer if selection acting on specific functions and pathways differed along the *Squalius* phylogeny, i.e. if the mode of selection changed along particular branches of the species tree. We used the R package RDAVIDWebService (Fresno and Fernandez, 2013) to test if the pOGs of Dataset C and Dataset D were enriched for specific functional annotations. RDAVIDWebService is an R interface of the DAVID Bioinformatics Resources v6.8 (Huang et al., 2009), which allows users to use the functionalities of the DAVID website (https://david.ncifcrf.gov) with the replicability advantages of R. For each dataset we used the list of pOGs as our test list and used the list of all OGs tested by aBSREL as the background list. This way, we account for the effect of potential ascertainment bias on the initial selection of genes we were able to align. Then we used the getClusterTermReport() function of RDAVIDWebService to run the Functional Annotation Clustering tool from DAVID. This tool works slightly differently from conventional annotation enrichment analysis. It starts as a conventional annotation enrichment analysis by testing if there are functional annotations overrepresented on the list of genes we are testing. Next, the algorithm clusters redundant functional annotations based on their co-occurrence on the genes of our test list. Finally, an enrichment score for the cluster of redundant annotations is calculated as –log(gmean(pval)), where gmean is the geometric means of the enrichment p-values of the annotations that form a given annotation cluster (Huang et al., 2009). For each annotation cluster, we created a representative label based on the annotations found on that cluster, especially based on the top scoring ones. This approach increases the power of the enrichment analysis by taking in consideration the redundancy of the annotation databases. We ran getClusterTermReport of RDAVIDWebService with default settings, except for the annotation databases. Besides the default databases, we also included the following databases to increase the power of our analysis:*"goterm_bp_all", "goterm_mf_all", "goterm_mf_fat", "pir_summary", "sp_comment", "pubmed_id", "reactome_pathway", "pfam", "prosite", "up_tissue"*.

2.3.3 - Functional enrichment mapping on the phylogeny

In order to gain insights on which biological processes have been targeted by natural selection at different time points in the phylogeny of these West Iberian *Squalius*, we considered doing separate enrichment analysis for pOGs inferred to be under selection on each time point of the species tree. According to the authors of the method implemented in DAVID, a minimum of 100 genes is advised in the test set for the enrichment analysis to have enough statistical power (Huang et al., 2009). We were unable to meet this minimum number of pOGs on the subset of pOGs under selection for each single branch of the phylogeny, so we used an alternative strategy. We did the functional enrichment analysis described above using as test

list all the pOGs, regardless of the branch of the phylogeny where we found the signal of positive selection. Then, for each of the top 10 annotation clusters, we identified the proportion of pOGs with evidence of selection on different branches of the species tree (species phylogeny). Note that when a pOG showed signatures of positive selection in a branch incompatible with the species tree we assumed it occurred in the closest ancestral species. For instance, in the case of selection on the ancestral branch of *S. carolitertii* and *S. torgalensis* in a given gene tree, we assumed that it was due to incomplete lineage sorting and that selection happened in the ancestral population of all *Squalius* species. Afterwards, we tested if there was an interaction between functional annotation cluster and the species tree branch where we found signals of positive selection, using a Chi-Squared test ($p\text{-val} < 0.05$). This was done in order to test if pOGs related to certain functional annotation showed a different pattern of selection across the phylogeny of the *Squalius* species, compared to the general pattern of selection for all the pOGs.

2.4 – Target genes related to temperature response and circadian rhythm

We selected a group of genes related to temperature and circadian rhythm found in previous works on western Iberian *Squalius* (Supplementary Table 2.1 and Supplementary Table 2.2) and searched for them on the OGs of Dataset C and D. This was done with a different goal for each set of genes. In the case of the genes related to temperature response (Supplementary Table 2.1), these have been proposed to be involved on adaptation to temperature in *S. torgalensis* (Jesus et al., 2016, 2017) but they were never tested for selection, therefore we wanted to know if in our datasets these genes showed signatures of selection supporting this hypothesis. On the other side, the genes related to circadian rhythm (Supplementary Table 2.2) have already been shown to exhibit signatures of positive selection (Moreno, J., 2018), thus, we used them as a way to test the sensibility of our Pipeline 3. We searched for those target genes in dataset C and D. If present, we checked whether they showed signatures of positive selection.

2.5 - Ethics statement

The specimens of *S. aradensis* were captured under license 421/2017/CAPT issued by Portuguese authority for Conservation of endangered species (ICNF - Instituto da Conservação da Natureza e das Florestas).

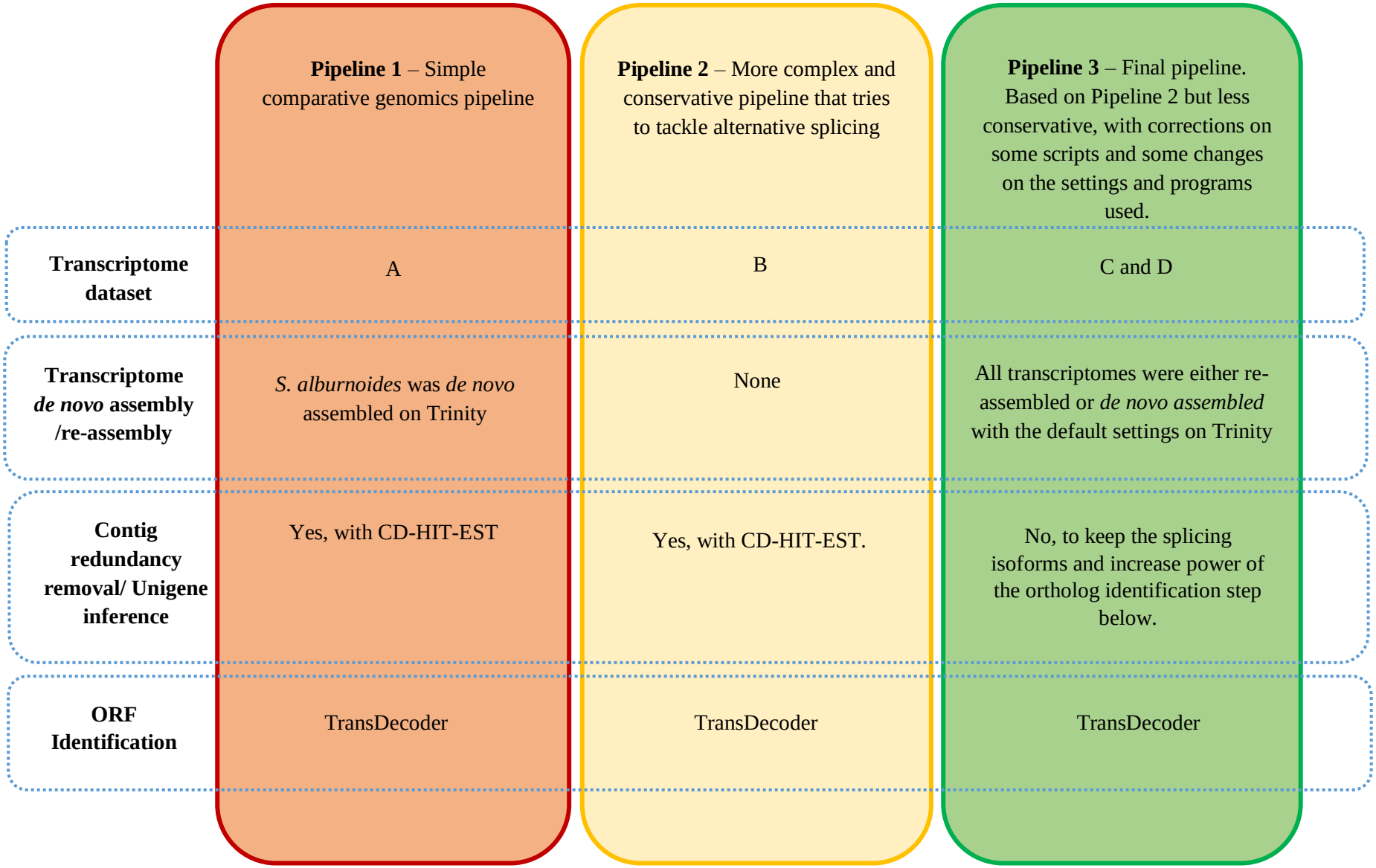


Figure 2.1 - Summary of the main differences between the three bioinformatic pipelines developed during this study to get OG alignments and test them for selection.

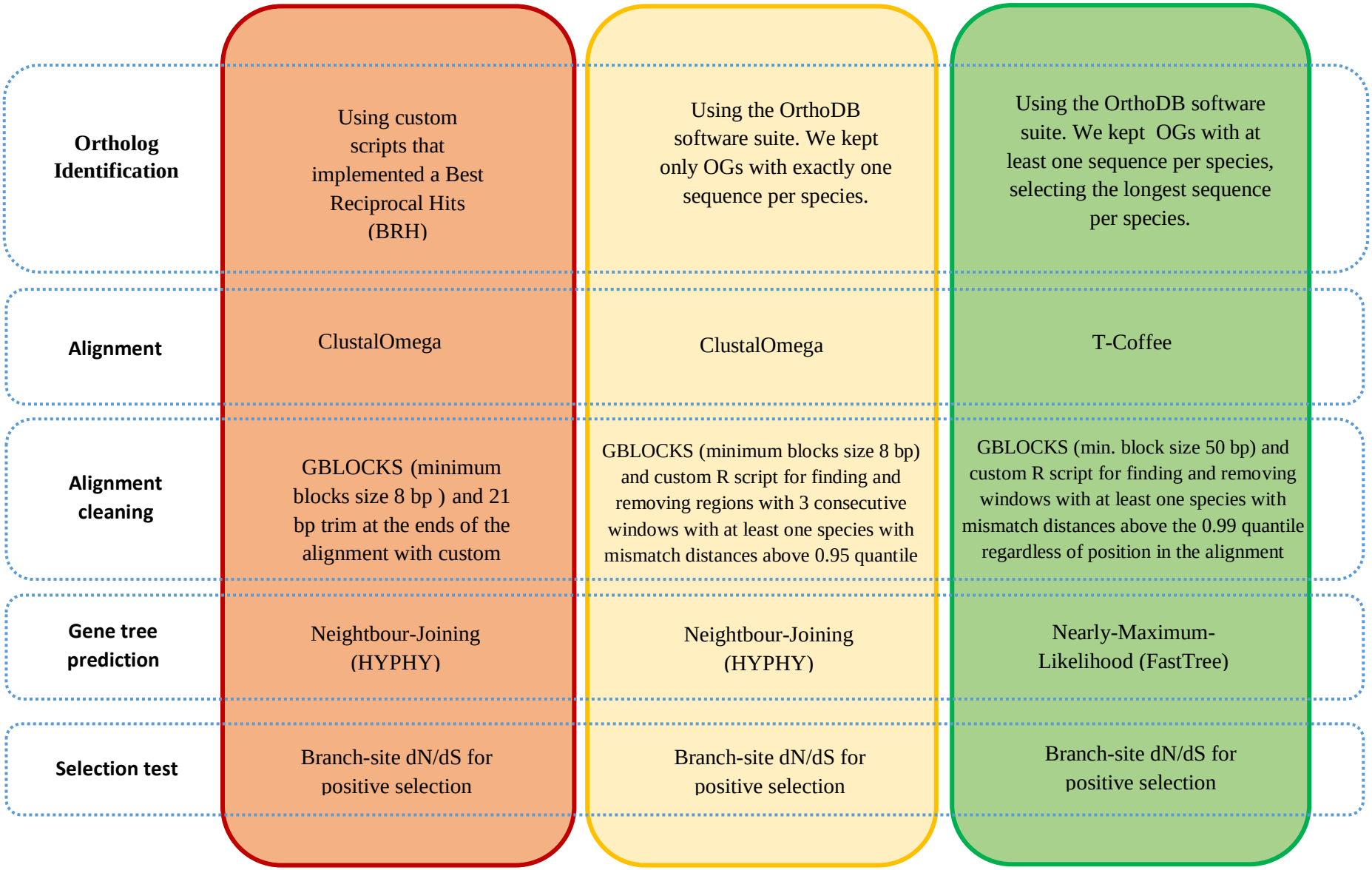


Figure 2.1 (Cont.) - Summary of the main differences between the three bioinformatic pipelines developed during this study to get OG alignments and test them for selection.

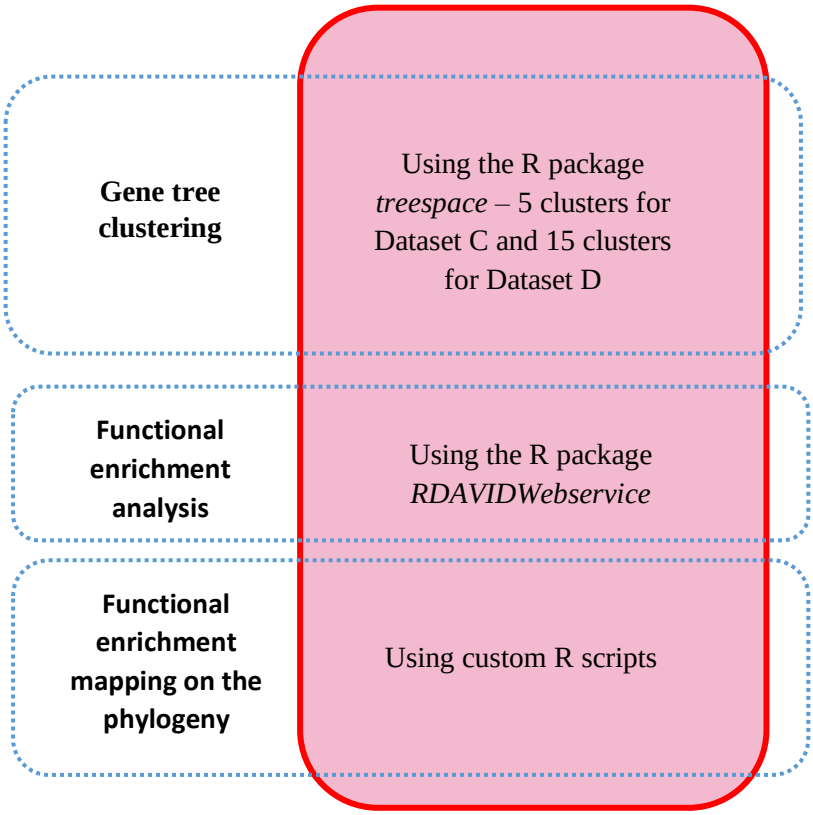


Figure 2.2 - Downstream analysis with the results of Dataset C and D

3. Results

3.1 - Transcriptome sequencing and assembly

To study the signatures of positive selection on the transcriptome of all the Portuguese *Squalius* species, we used both transcriptomic data from previous studies and new data to cover all the species and most of their distribution area in Portugal. We sequenced for the first time the transcriptomes of *S. aradensis* and *S. pyrenaicus* from the Tagus basin, which were species and populations without any transcriptomic data available before. We also sequenced for the first time a transcriptome for the *S. pyrenaicus* of the Guadiana basin obtained from muscle tissue. For these three new transcriptomes the number of raw reads obtained varied from 67,875,496 to 72,400,734 reads and the percentage of clean reads varied between 91.66% to 97.31% (Table 3.1). Even though the *S. aradensis* read library was the one with more raw reads (77,400,734), it was also the one with more reads lost at the cleaning step, with only 91.66% of its original reads kept, suggesting a lower quality for the *S. aradensis* library.

Table 1.1 – Raw and cleaned reads metrics for the three transcriptomes sequenced in this study. Gb stands for giga bases (1 Gb = 1 000 000 000 base pairs)

Transcriptome	Tissue	Total raw reads	Total clean reads	Total clean bases (Gb)	Percentage of clean reads (%)
S. pyrenaicus (Tagus)	Muscle	67 875 496	65 844 382	6.58	97.01
S. pyrenaicus (Guadiana)	Muscle	67 875 780	66 048 410	6.60	97.31
S. aradensis	Brain	72 400 734	66 362 238	6.04	91.66

Regarding the assemblies contig metrics (Table 3.2), which were computed for all transcriptomes included in our datasets after ORF identification, the final number of contigs on the assemblies varied from 127,405 on the preliminary *S. pyrenaicus* (SpyrenTP) assembly to 659,364 contigs on the published *L. burdigalensis* (LburdiO) assembly. The smallest contig size for all assemblies was 201 bp as expected since the minimum contig size parameter on Trinity was set to 200 bp. Mean contig length ranged from 517.9 bp in the *S. aradensis* (Sarad) assembly to 1,285.5 bp on the published *L. burdigalensis* assembly (LburdiO). However, note that the *S. aradensis* (Sarad) assembly seemed to be an outlier, as the second shortest mean contig length was of 769.2 bp (Table 3.2). For the N50 length (which represents the length at which all contigs with that length or greater include at least 50% of all the nucleotides bases on the assembly) we saw a similar pattern. N50 ranged from 655 bp on the *S. aradensis* assembly (Sarad) to 2,616 bp for the published *L.*

burdigalensis assembly (LburdiO). *S. aradensis* (Sarad) also seemed an outlier as the second shortest value was of 1,340 bp. Finally, GC content ratio varied between 0.42 also in *S. aradensis* (Sarad) to 0.46 on several other assemblies (Table 3.2). In general, in terms of number of sites and contig length distribution, the results showed a lower quality for the *S. aradensis* assembly (Sarad) and a relatively higher quality across all the other assemblies, with the published *L. burdigalensis* assembly (LburdiO) seemingly the one with the best quality of all.

The results for the completeness of the transcriptomes assemblies showed a similar pattern, indicating lower quality for the *S. aradensis* assembly (Sarad) and higher and more similar quality for the other species. Of the 4,584 Actinopterygii single copy orthologs that BUSCO searched on the assemblies, only 37.7 % were present on the *S. aradensis* assembly (Sarad), whereas between 74.5% and 92.3% were present on all the other assemblies (Table 3.3). Conversely, the *S. aradensis* assembly (Sarad) showed 20.3% and 42.0% of fragmented and missing orthologs, respectively, while all the other assemblies exhibited lower values ranging from 4.4 % to 14.5% of fragmented orthologs and from 2.5% to 14.6% of missing orthologs. Taken together with the results mentioned above, these results indicate that our *S. aradensis* transcriptome assembly was more fragmented than the rest of the assemblies. In a comparative study like this, having a transcriptome assembly with relatively more fragmented sequences can have a strong impact on the results, due to the requirement of generating alignments with orthologous sequence from every species. Thus for the downstream analyses, we repeated all the analysis of Pipeline 3 with two datasets, one without and another with *S. aradensis* (Datasets C and D, respectively).

The proportion of conserved orthologs classified as "duplicated" by BUSCO was very high on all our assemblies, ranging from 13.2% to 81.8% of the 4,584 Actinopterygii universal single-copy orthologs (Table 3.3 and Figure 3.1). This high duplicated proportion was expected for our RNA-seq dataset due to the presence of alternative splicing on the transcriptomes, with several isoforms for each transcript. Usually these isoforms are removed before the BUSCO analysis, however removing them at this stage decreased the power of OrthoDB for finding orthologous groups. Thus, the high level of duplicated conserved orthologs given by BUSCO could be interpreted as a proxy of the level of isoform representation on each transcriptome assembly. This is supported by the fact that the assembly with better quality, the original *L. burdigalensis* assembly (LburdiO), exhibited the greater proportion of duplicated orthologs, and inversely the *S. aradensis* assembly (Sarad), the one with lesser quality, exhibited the lesser proportion of duplicated orthologs.

Table 3.2 - Contig summary statistics for all the transcriptomes assemblies used on this study. “bp” stands for base pair, and Mb stands for mega bases (1 Mb = 1 000 000 base pairs). All assemblies with a name ending in “O” are published transcriptomes. See Table 2.1 for details on transcriptome assembly codes.

Transcriptome assembly	Dataset	Number of sequences	Smallest (bp)	Largest (bp)	Total bases (Mb)	Mean length (bp)	N90 (bp)	N50 (bp)	N10 (bp)	GC content ratio
ScarolO	A and B	145 975	201	17 990	117.07	801.96	302	1454	4622	0.45
ScarolR	C e D	195 069	201	15 864	183.19	939.09	341	1 773	4 898	0.45
SpyrenTP	C	127 405	201	15 046	123.19	966.92	355	1 779	4 925	0.46
SpyrenTF	D	178 663	201	15 378	165.83	928.18	352	1 616	4 027	0.45
SpyrenGO	A, B and C	252 870	201	16 728	235.49	931.26	323	1 844	5 222	0.46
SpyrenGR	C	195 069	201	15 864	183.18	939.09	341	1 773	4 898	0.45
SpyrenGD	D	280 573	201	23 348	257.01	916.04	325	1 754	4 664	0.45
StorgalO	A and B	137 303	201	19 053	105.61	769.21	296	1 340	4 068	0.45
StorgalR	C and D	185 569	201	18 619	170.88	920.86	338	1 695	4 423	0.45
Sarad	D	165 703	201	8 601	85.82	517.94	244	655	2 018	0.42
Salbur	A	243 338	201	25251	242.05	994.69	342	1972	5215	0.46
LburdiO	A and B	659 364	201	20 930	847.63	1285.53	470	2 616	6654	0.43
LburdiR	C and D	373 385	201	20 288	378.26	1013.06	353	2 079	5 648	0.43

Table 3.3 - Completeness results of the transcriptome assemblies used in this study as calculated by BUSCO using a benchmark of 4,584 universal single copy genes on Actinopterygii. Values represent the percentage of those 4,584 conserved orthologs that were found in each of the categories of BUSCO. The "Single-copy" and "Duplicated" categories are both included on the "Complete" category. See Table 2.1 for details on transcriptome assembly codes.

Transcriptome	Dataset	Complete (%)	Single-copy (%)	Duplicated (%)	Fragmented (%)	Missing (%)
ScarolO	A and B	79.1	57.1	22.0	9.5	11.4
ScarolR	C and D	80.9	42.9	38.0	9.2	9.9
SpyrenTF	C	74.9	46.5	28.4	10.5	14.6
SpyrenTF	D	74.5	37.8	36.6	14.5	11.0
SpyrenGO	A and B	88.4	52.7	35.7	4.6	7.0
SpyrenGR	C	87.8	40.2	47.6	7.0	5.2
SpyrenGD	D	82.7	35.0	47.7	11.1	6.2
StorgalO	A and B	79.1	54.7	24.3	9.7	11.3
StorgalR	C and D	79.3	40.1	39.2	10.2	10.5
Sarad	D	37.7	24.5	13.2	20.3	42.0
Salbur	A	92.3	43.1	49.2	5.2	2.5
LburdiO	A and B	90.0	8.2	81.8	4.4	5.7
LburdiR	C and D	86.8	32.0	54.8	5.5	7.7

Finally, after assessing the quality of the transcriptome assemblies, we used Transdecoder to predict the sequence's open reading frames (ORF's) for all the assemblies. We then created a coding sequence database for each species, containing the longest ORF of each contig on the transcriptome assembly. The percentage of contigs for which Transdecoder was able to predict an ORF varied from 23.6 % on *S. aradensis* to 41.1 % on preliminary Tagus *S. pyrenaicus* assembly (SpyrenTP) (Table 3.4).

3.2 Ortholog groups (OGs) and multispecies alignments

We developed and tested three different bioinformatic pipelines to identify and create good quality alignments of the orthologous groups (clusters of orthologous sequences) between the transcriptomes of our species. Since the datasets were not the same between pipelines because the work was done sequentially (see methods), the results of the different pipelines cannot be directly compared. Still, we could evaluate the relative effect of each pipeline on the resulting number of orthologous groups (OG), number of cleaned alignments and number of OGs and genes with evidence of positive selection (Figure 3.2 and Table 3.5). We found that the number of OG identified varied considerably depending on the pipeline and dataset we used, ranging from 475 on Dataset B of Pipeline 2 to 13,525 in Dataset C with Pipeline 3 (Figure 3.2 and Table 3.5). We also found that in all the pipelines there were certain OGs where the alignment algorithm

failed. In Pipeline 1, this happened only for 1 of the 10,767 OGs, but both in Pipeline 2 and Pipeline 3 we found more cases. From 475 OGs on Dataset B of Pipeline 2, we obtained 419 alignments (88.2%); from 13,525 OGs on Dataset C of Pipeline 3 we obtained 13,300 alignments (98.3 %), and from Dataset D of Pipeline 3 from 9,604 OGs we obtained 9,464 alignments (98.5 %) (Table 3.5). Even though the Datasets used were different, these results suggest that Pipeline 3 is the less stringent for detecting OGs, while Pipeline 2 is the most conservative, resulting in a very reduced number, whereas Pipeline 1 is intermediate.

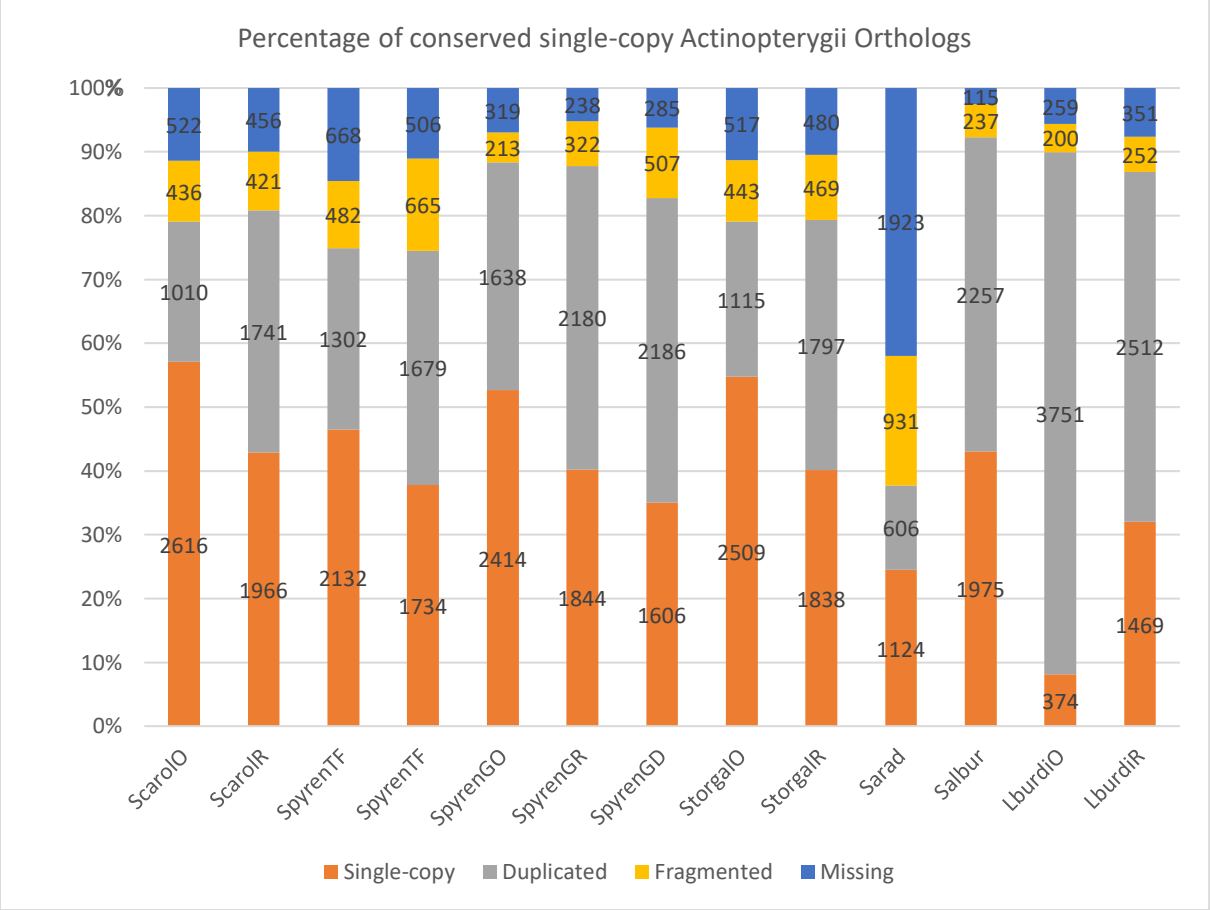


Figure 3.1 - BUSCO results for the transcriptome assemblies used in this work. Each column represents the results for one transcriptome assembly and each color represents how many of the 4584 conserved orthologs fell on each one of the categories of BUSCO, except the Completed category which is the sum of the proportions of the Single-Copy and Duplicated categories. Numbers on the colored regions describe the exact number of conserved orthologs found on the category represented by that color. See Table 2.1 for details on transcriptome assembly codes.

Table 3.4 – Number and proportion of open reading frames (ORF's) predicted by Transdecoder for each transcriptome assembly. See Table 2.1 for details on transcriptome assembly codes.

Transcriptome	Dataset	Number of contigs	Number of ORF's	Percentage of contigs with ORF's (%)
ScarolO	A and B	145 975	48 439	33.2
ScarolR	C and D	195 069	72 580	37.2
StorgalO	A and B	137 303	45 984	33.5
StorgalR	C and D	185 569	70 358	37.9
SpyrenTP	C	127 405	52 341	41.1
SpyrenTF	D	178 663	73 204	41.0
SpyrenGO	A and B	181 473	58 147	32.0
SpyrenGR	B	252 870	87 080	34.4
SpyrenGD	A	280 573	95 958	34.2
Sarad	D	165 703	39 077	23.6
Salbur	A			
LburdiO	A and B	659 364	24 4075	37.0
LburdiR	C and D	373 385	113 686	30.4
Danio rerio (CDS)	A, B, C and D	-	46 260	-

Table 3.5 - Summary of the results of several steps of the bioinformatic pipeline from the ortholog identification until the selection tests. The proportions indicated refer to the proportion in relation to previous step of the pipeline.

Step	Dataset A (Pipeline 1)	Dataset B (Pipeline 2)	Dataset C (Pipeline 3)	Dataset D (Pipeline 3)
Number of OG's found by ORTHODB	10 767	475	13 525	9 605
Proportion of raw alignments	1.000	0.882	0.983	0.985
Proportion of cleaned alignments	1.000	0.914	0.923	0.804
Proportion of OG’s succesfully tested for selection by aBSREL	0.910	0.943	1.000	1.000
Proportion of OGs with signatures of positive selection (pOGs)	0.133	0.053	0.020	0.014

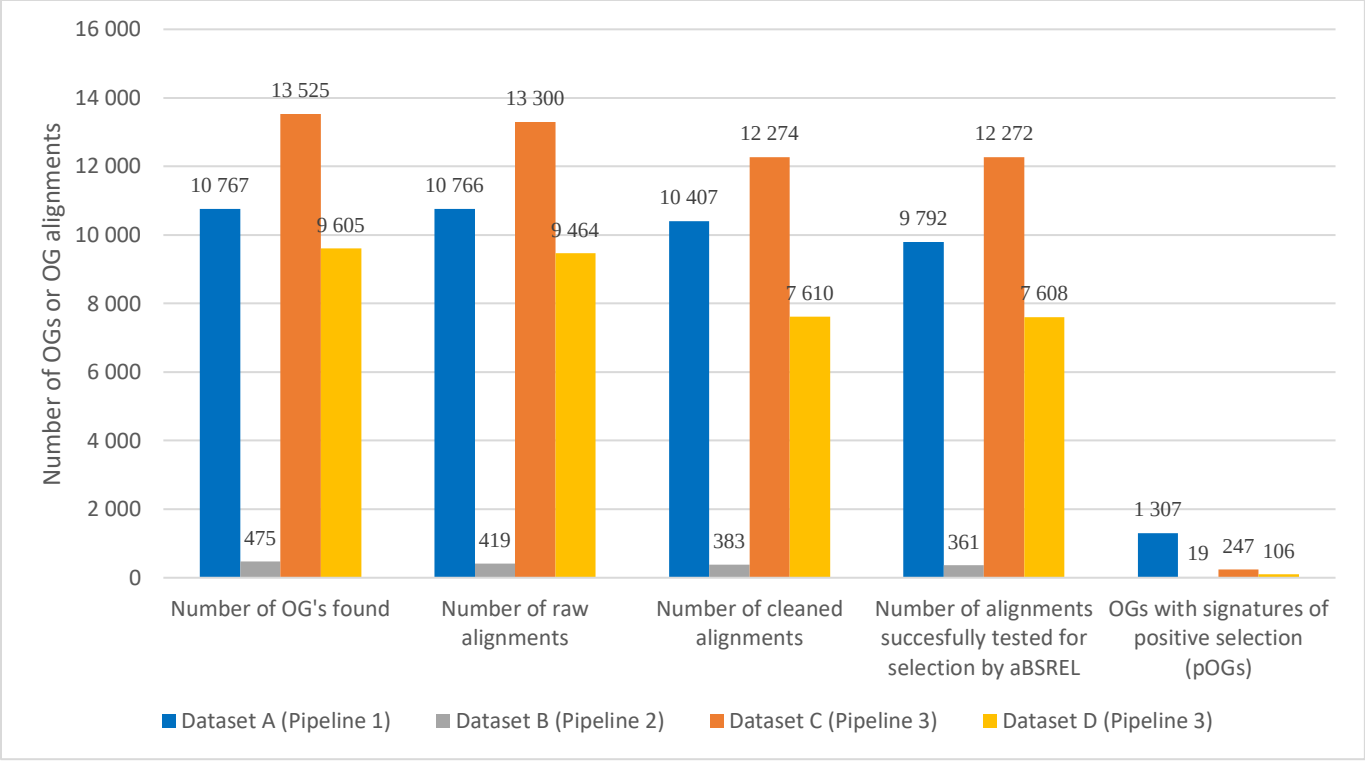


Figure 3.2 - Number of orthologs throughout the different steps of the bioinformatic pipeline from the ortholog groups identification until the selection tests. Blue represents the values for Dataset B and orange the values for Dataset A. Numbers above the bars represent the number of OGs or OGs alignments.

For all datasets, the alignment cleaning step reduced slightly further the resulting number of OGs. For Dataset A using Pipeline 1 we obtained 10,407 cleaned alignments (96.7% of the raw alignments), for Dataset B using Pipeline 2, we obtained 383 clean alignments (91.4% of the raw alignments) and for Datasets C and D using Pipeline 3, we obtained 12,274 and 7,610 cleaned alignments (92.3% and 80.4% of the raw alignments), respectively (Figure 3.2 and Table 3.5). This suggests that this step had a similar effect across pipelines, but that Pipeline 3 was more stringent as it discarded more alignments. For Pipeline 3 we also calculated the mean length of the cleaned alignments, which was 1,014.5 bp for Dataset C, and 600.7 bp for Dataset D (Figure 3.3).

The lesser number of OGs and shorter alignment length on Dataset D were expected because 1) this dataset had one extra species (*S. aradensis*) relatively to Dataset C, which decreases the probability of finding OGs and regions on the alignments with sequence for all species; and 2) the effect of the greater fragmentation of the *S. aradensis* assembly (Sarad). However, despite these two factors, Dataset D only had 22.3% less cleaned alignments than Dataset C, and the mean of its alignments was 40.8% shorter than the mean of Dataset C. Thus, we were still able to obtain a dataset that was useful for comparative analyses.

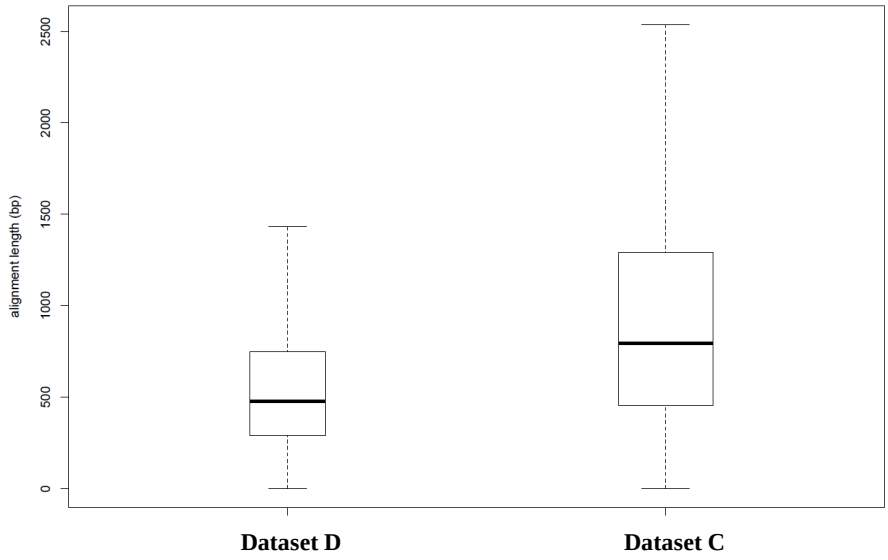


Figure 3.3 - Distribution of the alignment lengths on datasets C and D. Upper and Lower whiskers extend to the extremes of the distribution.

3.3 – Gene trees across the transcriptome

We calculated inferred gene trees for all the cleaned ortholog alignments on all our four datasets, which were used to perform the positive selection test. For Dataset C and D, we further analysed their gene trees (12,274 and 7,608, respectively). We grouped the gene trees estimated for each OG into gene tree clusters (GTCs) (5 GTCs for Dataset C and 15 for Dataset D, corresponding to the number of possible permutations between the ingroup *Squalius* species on each dataset). For each cluster we estimated the median topology to gain insights of the most supported gene trees across the transcriptome (Figure 3.4 and Table 3.6).

For Dataset C, 39.1% of genes were clustered into the topology that agrees with previously inferred species tree and forms a clade joining *S. carolitertii* and *S. pyrenaicus*, but showing Tagus *S. pyrenaicus* closer to *S. carolitertii* (C1 in Figure 3.4 - I). The second most common topology, with 28.0% of gene trees, supported a polytomy of all *Squalius* species. We did not find support for the clustering the two *S. pyrenaicus*, but found many polytomic clusters together with *S. carolitertii* (C3 in Figure 3.4 - I). Interestingly, for cluster C5 the gene tree show a separation of species according to north-south (Atlantic climate-type vs Mediterranean climate-type), and hence the gene trees at those genes could result from the action of selection, namely convergence within species inhabiting the same climate-types. Indeed, incongruences between gene trees and the species tree topology can be due to the action of selection (Rosenberg and Nordborg, 2002) but can be also due to ancestral polymorphism (incomplete lineage sorting).

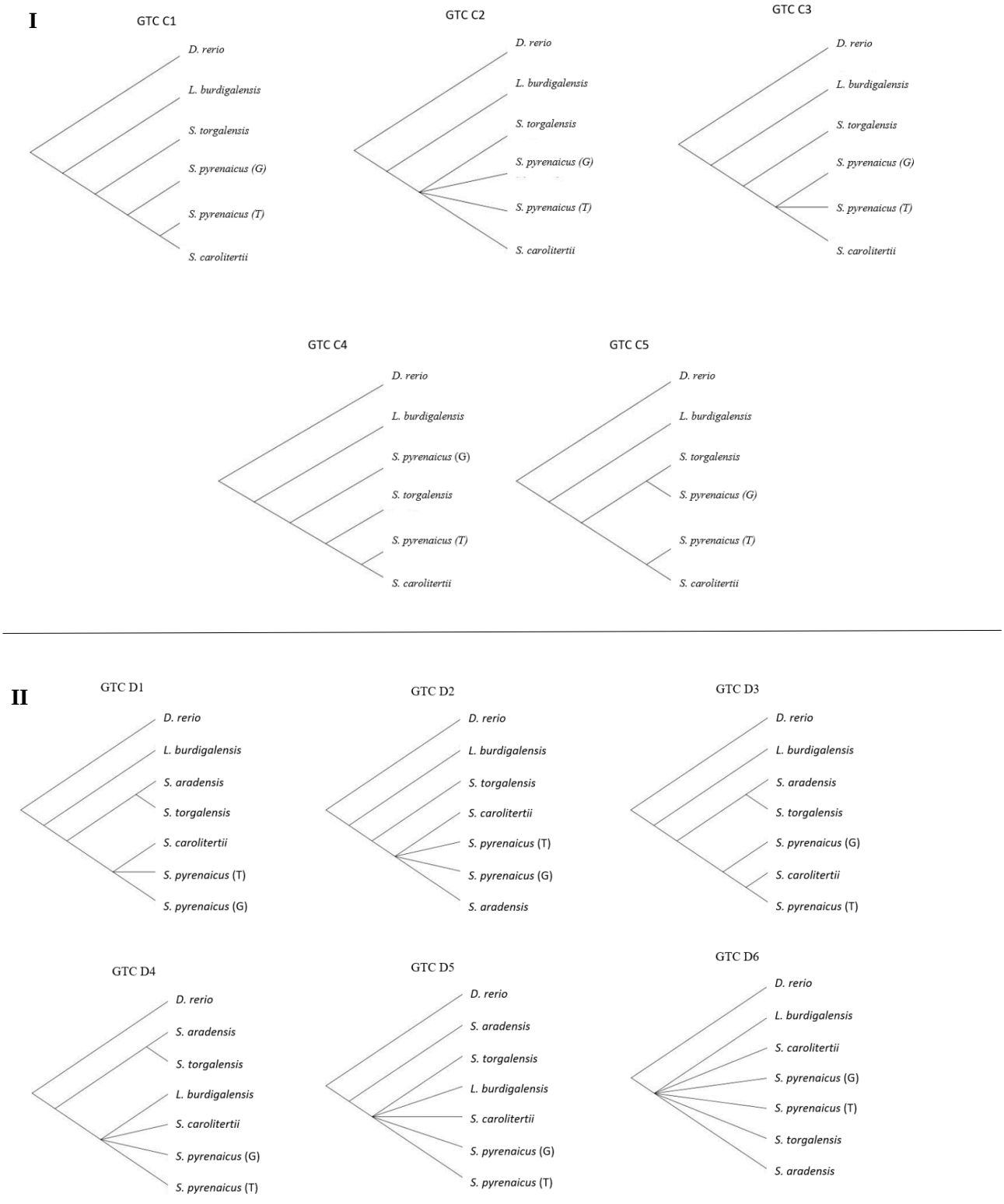


Figure 3.4 - Median topology of the gene tree clusters (GTC) on Pipeline 3 for: I) all the GTC's on Dataset C and II) the six top GTC's on Dataset D. S. pyrenaicus (G) refers to the Guadiana population of this species and S. pyrenaicus (T) refers to the Tagus population of this species.

The fact that we found most genes to fall within two clusters (C1 and C2, ~70%) in agreement with the species tree, and that the 3 remaining clusters (C3-C5, ~30%) have similar proportions (10-12%) indicate that the incongruences between the gene trees and species tree topologies are most likely due to ancestral polymorphism and not due to a major effect of positive selection driving towards a particular topology.

Regarding Dataset D, the most common topology with 27.0% of genes also matched the previously inferred species tree (C1 in Figure 3.4 and Table 3.6), distinguishing two main lineages of *Squalius*: (i) *S. carolitertii* and *S. pyrenaicus* clade, and (ii) *S. aradensis* and *S. torgalensis* clade. The other top 5 topologies had a similar proportion of genes from 7-9%, in agreement with what is expected due to neutral ancestral polymorphism, i.e. a similar value for different topologies incongruent with the species tree. However, most of the gene tree clusters revealed polytomic median trees, with some even clustering the *L. burdigalensis* outgroup together with the ingroup species (Figure 3.4.II and Supplementary Figure 3.1). Even so, only 5 of the 15 GTCs had median trees with polytomies that went beyond the clade of *S. carolitertii* and the two *S. pyrenaicus* populations. Both D1 and D3 clusters matched the species tree except for a polytomy on that clade on D1, indicating that for most genes there was still enough information to infer relationships between species.

Table 3.6 - Frequency and proportion of gene trees in each of the gene tree clusters (GTCs) for Dataset C and D.

Dataset	GTC	Frequency	Proportion
C	C1	4805	0.39
	C2	3442	0.28
	C3	1455	0.12
	C4	1395	0.11
	C5	1177	0.10
	total	12274	1.00
D	D1	2040	0.27
	D2	709	0.09
	D3	657	0.09
	D4	549	0.07
	D5	520	0.07
	D6	516	0.07
	others	2620	0.34
	total	7611	1.00

3.4 – Orthologous groups with signatures of positive selection (pOG)

Using aBSREL we not only tested for signatures of positive selection on the alignments of the four datasets but also identified in which branches of the gene tree the selection signal was found. There were cases, however, in which aBSREL was unable to perform the tests. This happened mostly in Pipeline 1 and Pipeline 2, since in those two pipelines we used custom algorithm for removing stop codons that in some situations led to artificial frameshifts on the alignments. In Pipeline 3 we corrected this, and as a result aBSREL worked for almost all of the tested alignments (Supplementary Files 3.1 and 3.2). We found that aBSREL ran the tests successfully in 9,792 out of 10,766 cleaned alignments (91,0%) for dataset A using Pipeline 1 and for 361 out of the 383 cleaned alignments 361 alignments (94.3%) for dataset B using Pipeline 2. This suggests that the artificial frameshifts introduced in some situations affected only a small proportion of alignments (<9%). In contrast, with Pipeline 3 only 2 alignments in each dataset failed to be tested by aBSREL, i.e. less than 0.0002% (Table 3.5).

Regarding the number of OGs with signatures of selection (pOGs – positive orthologous groups) we found great differences between pipelines. We found the higher number of pOGs on Dataset A with Pipeline 1 – (1,307 pOGs, 13.3% of all tested OGs), and the lower number for Dataset B with Pipeline 2 (19 pOGs, 5.3% of all tested OGs, Figure 3.2). Note that even for the case with more pOGs Dataset A (Pipeline 1), our estimates suggest that only a small proportion of genes are under positive selection (<14%). However, for Dataset A we found several examples of misaligned regions. Figures 3.5 to 3.7 show three examples of clearly misaligned sequences that show evidence of positive selection. This suggests that some of the signatures of positive selection are due to misalignments. On Pipeline 3, where we corrected for misalignments, we found between 1.4% and 2.0% of the OGs with signatures of selection (Table 3.5).

3.4.1 A new approach for correcting misaligned regions

To control for alignment errors, we developed an R script that corrects for misalignments missed by GBLOCKS (Supplementary Folder 2.1). This approach was implemented on Pipeline 2 and 3. To validate this approach, we tested real and simulated alignments with and without misalignments and found that it worked well. Using this script which compares the average pairwise distance of each species against all the others along the alignment, detecting outliers regions with high pairwise distance (Figure 3.8), we were able to clean the alignments of misaligned regions left by GBLOCKS (Figures 3.5 to 3.7 and Supplementary Files 3.3 and 3.4) For this reason, this suggests that the approach we used in Pipeline 3 could remove misalignment problems that could eventually lead to false positives.

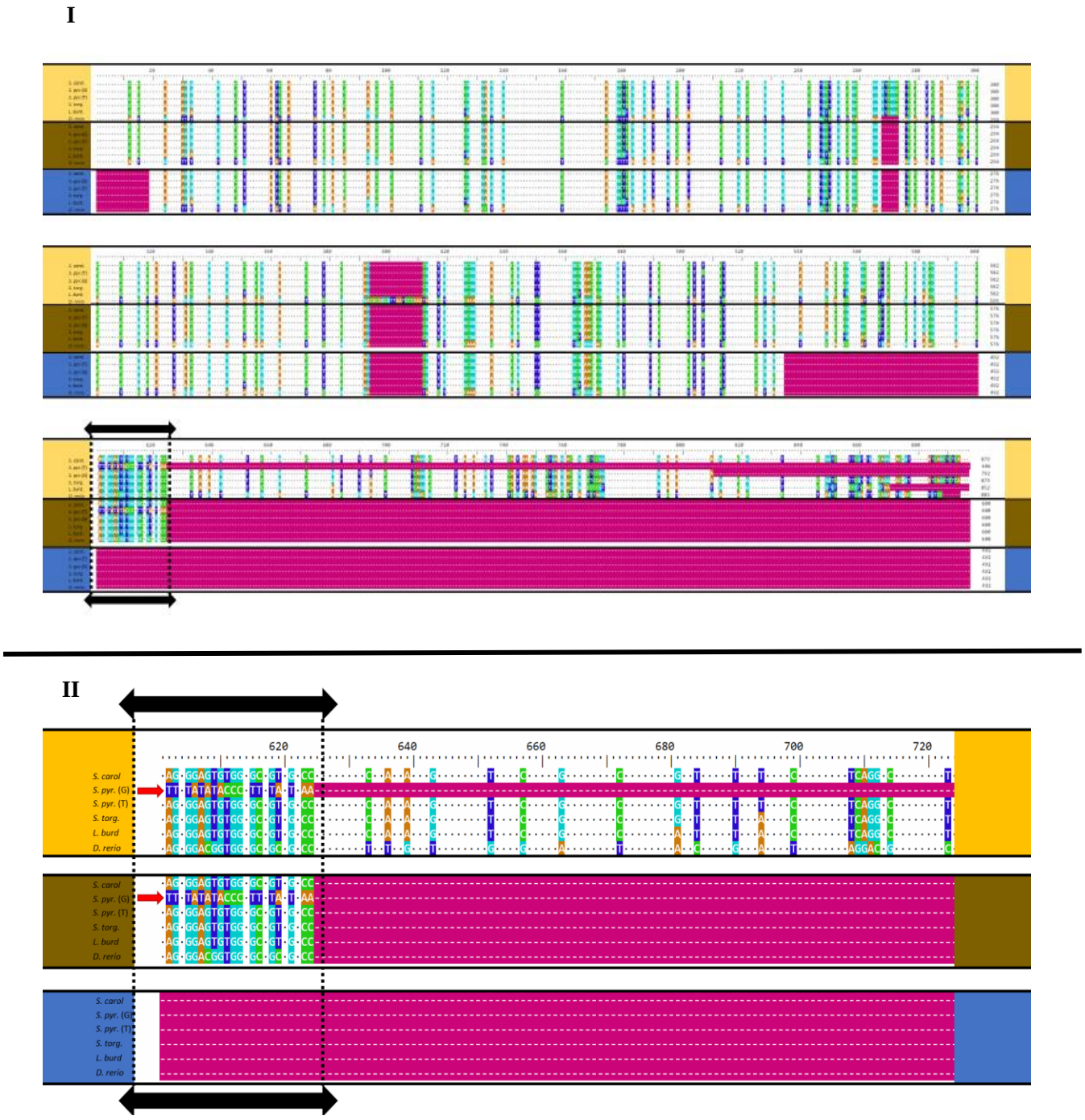


Figure 3.5 – First example of an alignment with a misaligned region in different steps of the cleaning process: raw (alignment delimited by orange boxes), after cleaning with gblocks (brown), and after cleaning with gblocks plus our custom script for removal of misaligned regions (blue). I) shows the whole alignment; II) Zooms in to the misaligned region. Red arrow points to the misaligned sequence. Loci on the alignment with a dot instead of the nucleotides represent conserved regions, loci with the coloured nucleotides represent either polymorphisms or gaps in that position. Pink means that sequences has no information on that loci – either because the sequence is incomplete, because it is a gap or because it was removed during the cleaning process. Black arrows represent the region of the alignment with the misaligned sequence not removed by GBLOCKS but removed by our script.

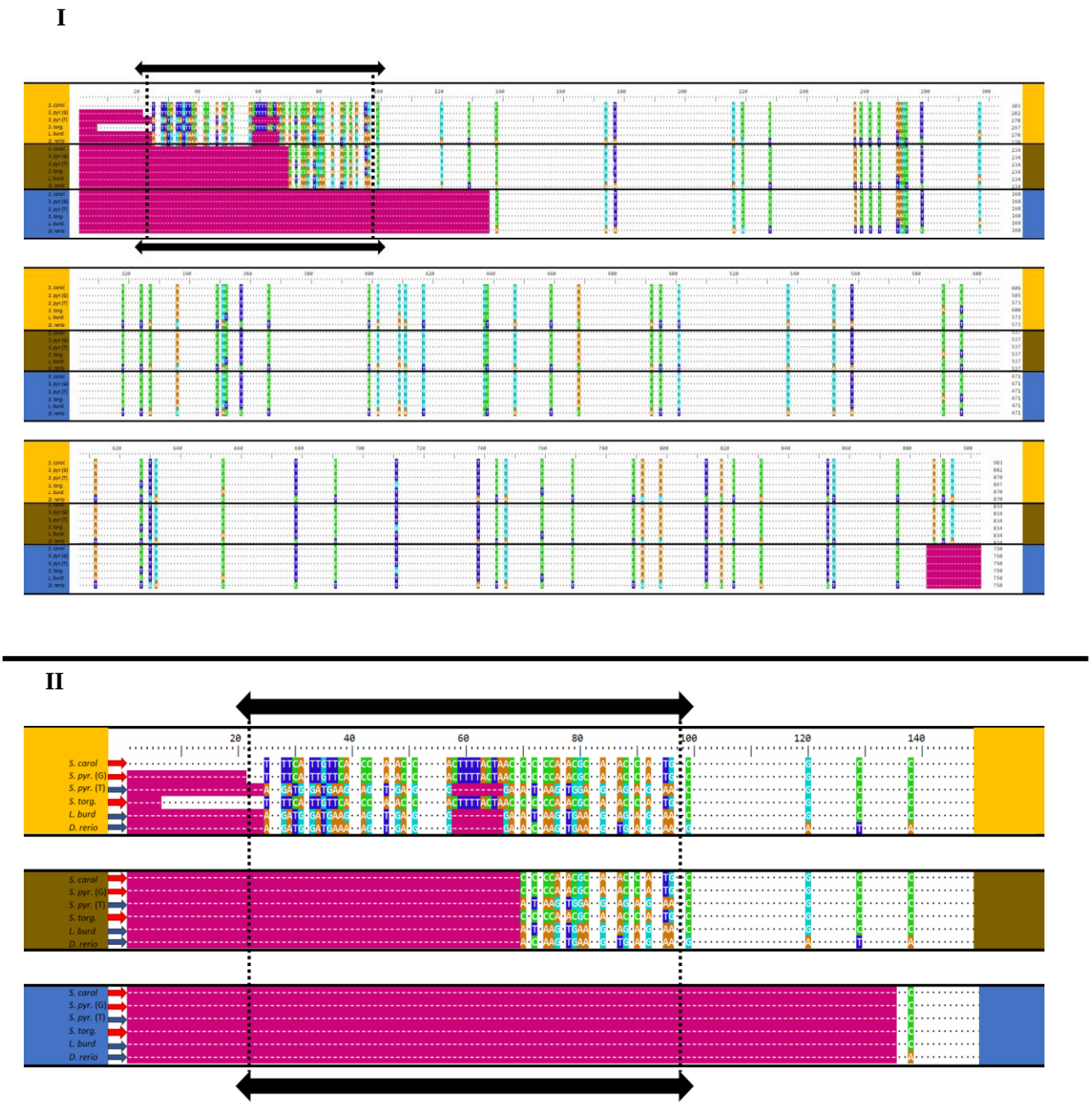
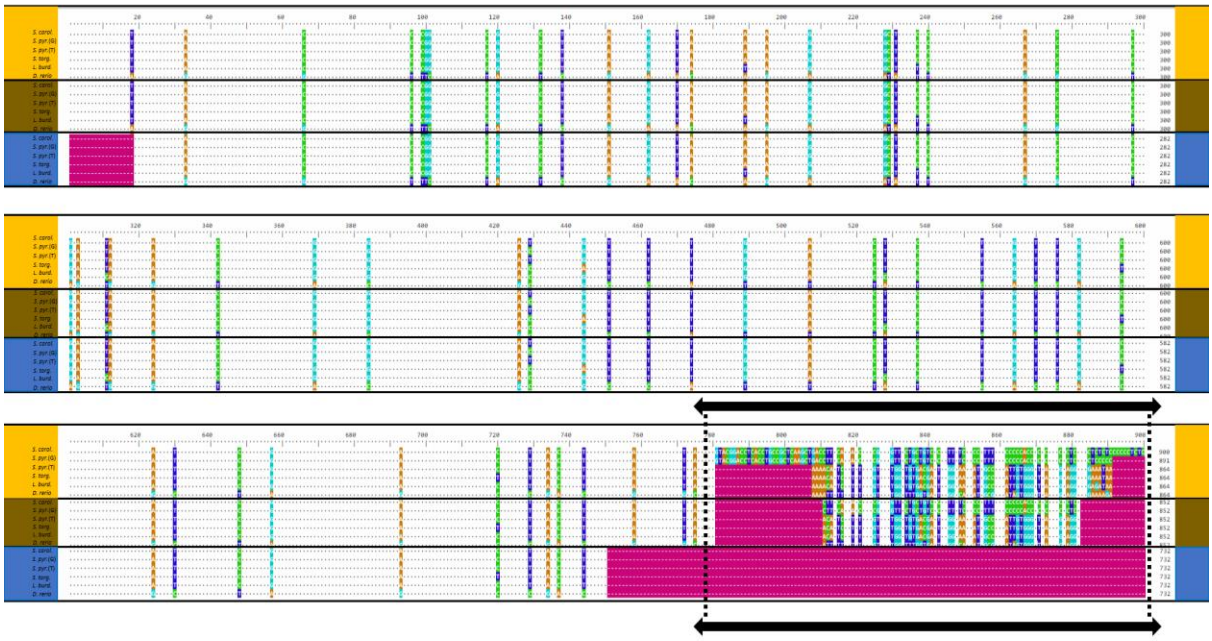


Figure 3.6 – Second example of an alignment with a misaligned region in different steps of the cleaning process: raw (alignment delimited by orange boxes), after cleaning with gblocks (brown), and after cleaning with gblocks plus our custom script for removal of misaligned regions (blue). I) shows the whole alignment; II) Zooms in to the misaligned region. In this case, the misaligned region clearly shows two different sets of sequences that most likely represent different exons of different isoforms. Red arrow point to the sequences potentially corresponding to one isoform and blue arrows to the sequences of the other isoform. Loci on the alignment with a dot instead of the nucleotides represent conserved regions, loci with the coloured nucleotides represent either polymorphisms or gaps in that position. Pink means that sequences has no information on that loci – either because the sequence is incomplete, because it is a gap or because it was removed during the cleaning process. Black arrows represent the region of the alignment with the misaligned sequence not removed by GBLOCKS but removed by our script.

I



II

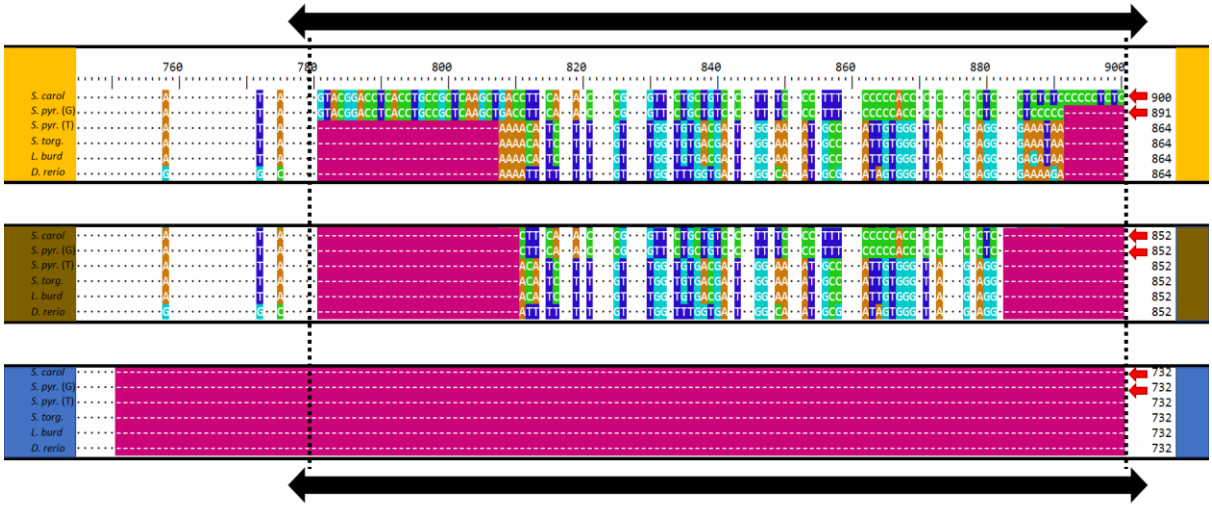


Figure 3.7 – Third example of an alignment with a misaligned region in different steps of the cleaning process: raw (alignment delimited by orange boxes), after cleaning with gblocks (brown), and after cleaning with gblocks plus our custom script for removal of misaligned regions (blue). I) shows the whole alignment; II) Zooms in to the misaligned region. In this case, the misaligned region clearly shows two different sets of sequences that most likely represent different exons of different isoforms. Red arrows point to the two sequences potentially corresponding to a different isoform from the rest. Loci on the alignment with a dot instead of the nucleotides represent conserved regions, loci with the coloured nucleotides represent either polymorphisms or gaps in that position. Pink means that sequences has no information on that loci – either because the sequence is incomplete, because it is a gap or because it was removed during the cleaning process. Black arrows represent the region of the alignment with the misaligned sequence not removed by GBLOCKS but removed by our script.

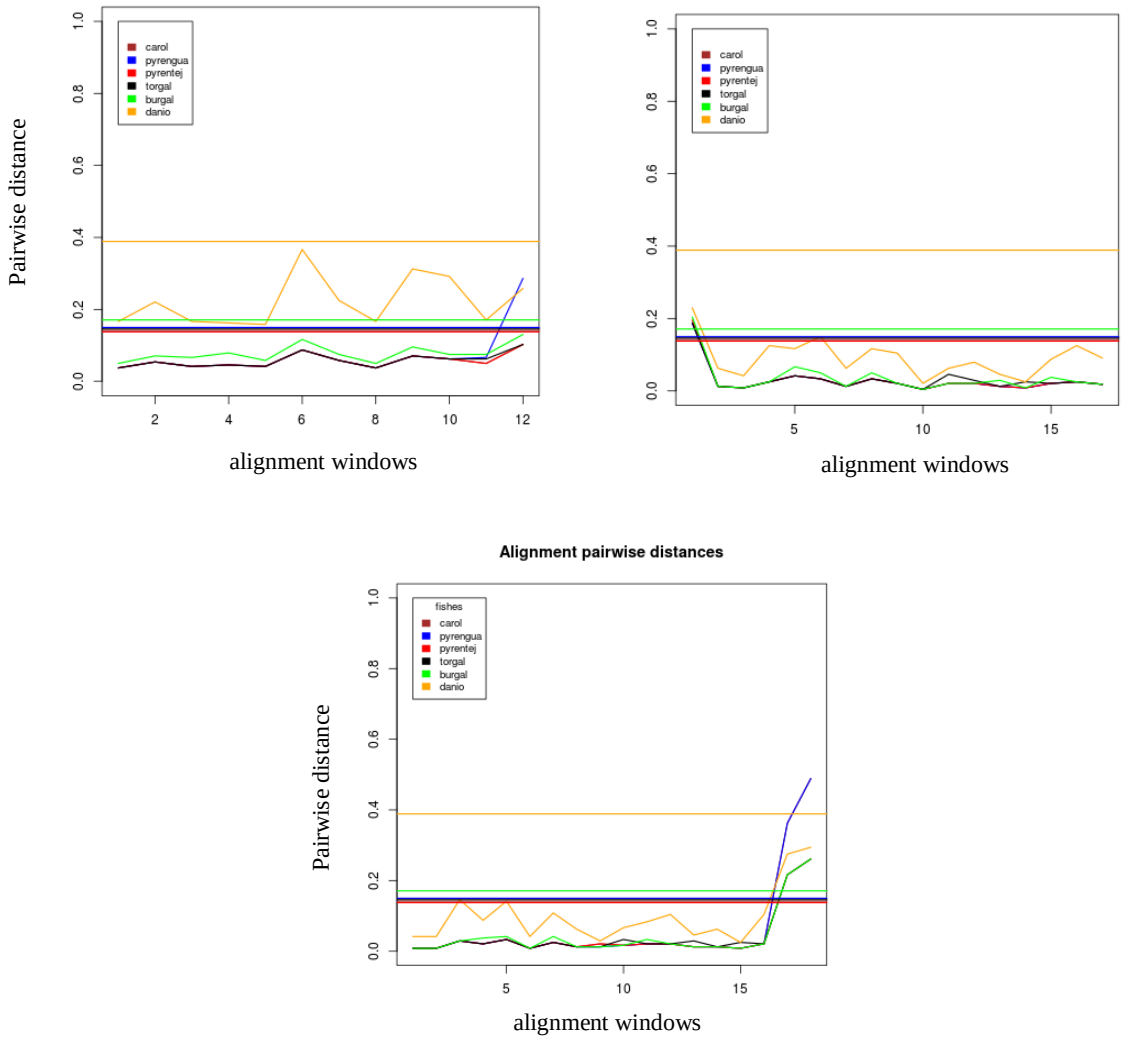


Figure 3.8 – Example of a plot of pairwise distances between the sequences of the different species on the three OG alignments showed above (From top-left to bottom: Figure 3.5, Figure 3.6 and Figure 3.7). Each colour represents a species as by the label on the plot. "Carol" – S. carolitertii; "pyrengua" – Guadiana population of S. pyrenaicus; "pyrentej" – Tagus population of S. pyrenaicus; "torgal" – S. torgalensis; "burgal" – L. burdigalensis; "danio" – D. rerio. Solid lines represent the average pairwise distance of the sequence of a given species against all the other species at each 48 bp window in the alignment. The pairwise distance was calculated as the proportion of mismatches on each sequence for a given window in the alignment. Horizontal lines represent the threshold of mismatches for each species. This threshold corresponds to the 95th quantile of the pairwise distances distribution for that species across all the windows of all the alignments, i.e. of all the orthologous groups. If one sequence crosses its respective threshold, the window on the alignment where that happens is removed from the alignment.

3.4.2 Patterns of positive selection across the western Iberian *Squalius* species tree

Of the 247 pOGs (2.0%) detected with Pipeline 3 on Dataset C (Figure 3.2), 12 had signatures of selection in more than one branch, bringing the total number of branches with signatures of selection on pOG to 261 (Supplementary File 3.5 and 3.7). To characterize the time (branch) on the phylogeny where most of positive selection occurred, we mapped the 261 branches inferred to be under selection into the inferred species tree. We found that all branches of the species tree had signatures of selection (Figure 3.9.I and Table 3.7), and no correlation between branch length and the strength of the signal (p-value = 0.10; Supplementary Figure 3.2.I). We also tested for differences between the proportion of OGs and the proportion of pOGs, but for most gene tree clusters this was not significant (Supplementary Figure 3.4.I). Many pOGs showed signs of positive selection in branches that were incongruent with the species tree, which given the proportion of the topologies (Figure 3.4-I and Table 3.6) we assumed was due to ancestral polymorphism. Hence, for those genes selection likely occurred in the ancestral population of the descendant species. We found that more than a third of the 89 pOGs (34.1%) mapped to selection on the ancestral of the *Squalius* clade and either *L burdigalensis* or *D. rerio* (Figure 3.9-I and Table 3.7, "Others"), suggesting that some of the beneficial mutations are old and were already segregating in the ancestral species. Still, we also found some genes to be under selection more recently (tips of the species tree), i.e. only on a given species branch, which varied from 5 pOGs (1.9 %) in *S. carolitertii* branch to 46 pOGs (17.6% in the Guadiana *S. pyrenaicus*). Interestingly, the number of pOGs under selection was very different between the two populations of *S. pyrenaicus.* While we found 46 pOGs for the Guadiana population, we only detected 6 for the Tagus population (2.3% of all significant branches on pOGs), which is very similar to the result of the *S. carolitertii* branch (5 pOGs). Moreover, we found that the branches of species inhabiting the sourthern river drainages have more pOG: 46 in *S. pyrenacus* from Guadiana, and 31 in *S. torgalensis*. This is in sharp contrast with the numbers detected for *S. carolitertii* and Tagus *S. pyrenaicus* species that live in more northern river drainages and seem to have less genes under positive selection (Figure 3.9 and Tables 3.7 and 3.8). However, we detected 36 genes under positive selection in the ancestor of Tagus *S. pyrenaicus* and *S. carolitertii*, which is similar to the values found in the southern basins. These results show that, at least in Dataset C, there seems to be a greater similarity between the patterns of past positive selection between the Tagus population of *S.pyrenaicus* and *S. carolitertii*, which share more genes under positive selection, than the Tagus and Guadiana *S. pyrenaicus* populations.

Regarding Dataset D, which included *S. aradensis*, we found 106 pOGs (1.4% of OGs) of which only one showed signatures of selection in more than one branch of the tree (Supplementary File 3.6 and 3.8). Just like for Dataset C we did not found a significant correlation between branch length and the strength of the positive selective signal (p-value = 0.59; Supplementary Figure 3.2.II), and most gene tree clusters did not

show a significant difference between the proportion of OGs and the proportion of pOGs (Supplementary Figure 3.4.II).

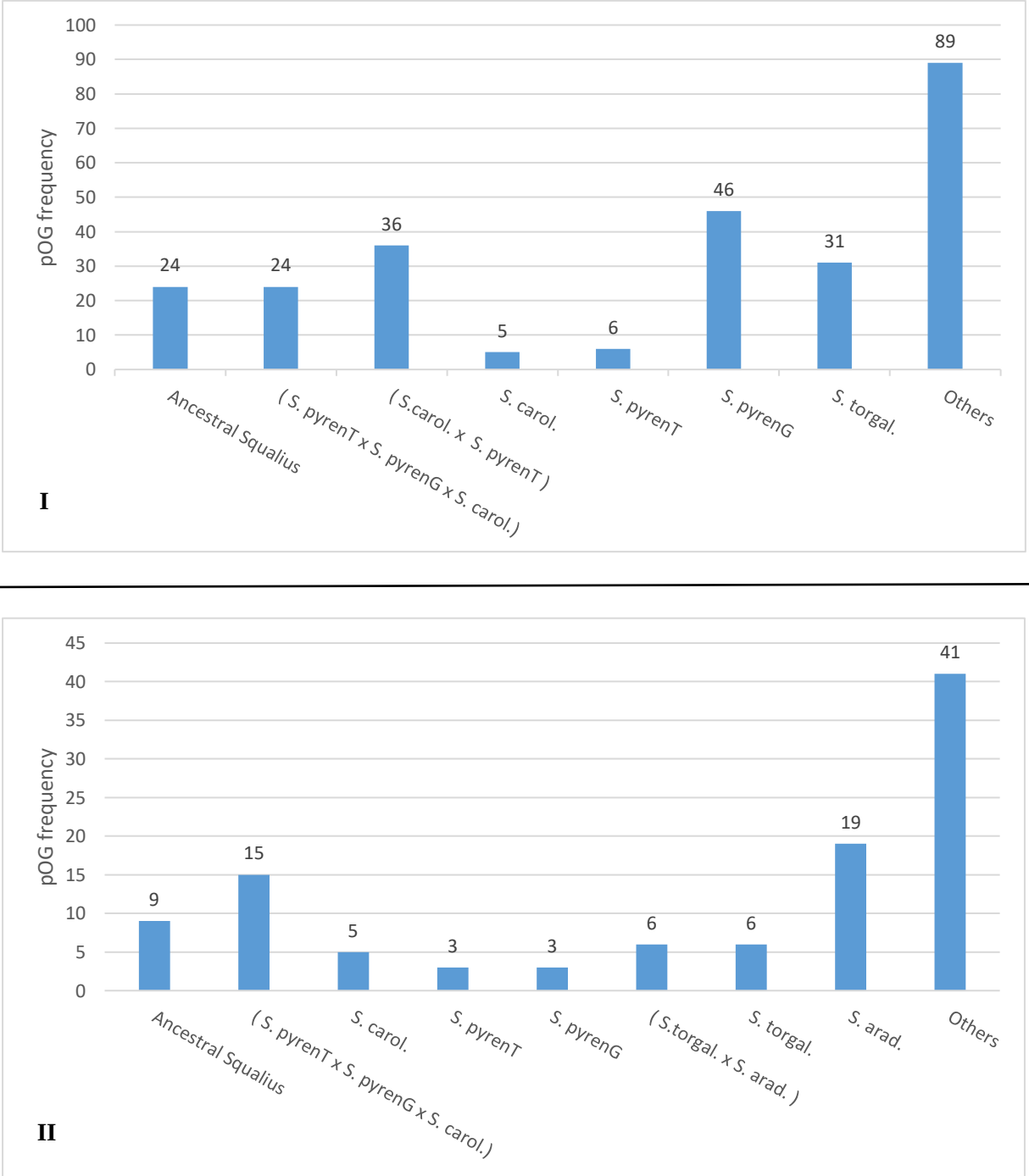


Figure 3.9 - Frequency of ortholog groups with signatures of positive selection (pOGs) on the branches of the inferred species tree (i.e the median topology of the most frequent gene tree cluster) of I) Dataset C and II) Dataset D. Branches with parentheses and several species names inside indicate the ancestral of those species. *S. carol.* – *S. carolitertii*, *S. pyrenT* – *S. pyrenaicus (Tagus)*, *S. pyrenG* – *S. pyrenaicus (Guadiana)*, *S. torgal* – *S. torgalensis*, *S. arad.* – *S. aradensis*, Ancestral *Squalius* – Ancestral branch of all *Squalius* species, Others – Ancestral of all the *Squalius* and one of the outgroup species.

Table 3.7 - Number and proportion of pOGs mapped on the different branches of the inferred species tree, which we assumed corresponded to the most supported *median gene tree topology* for Dataset C.

Branch	Number of pOGs	Proportion
Ancestral Squalius	24	0.09
(S. pyrenT x S. pyrenG x S. carol.)	24	0.09
(S.carol. x S. pyrenT)	36	0.14
S. carol.	5	0.02
S. pyrenT	6	0.02
S. pyrenG	46	0.18
S. torgal.	31	0.12
Others	89	0.34
Total	261	1.00

Table 3.8 - Number and proportion of pOGs mapped on the different branches of the inferred species tree, which we assumed corresponded to the most supported *median gene tree topology* for most supported GTC for Dataset D.

Branch	Number of pOGs	Proportion
Ancestral Squalius	9	0.08
(S. pyrenT x S. pyrenG x S. carol.)	15	0.14
S. carol.	5	0.05
S. pyrenT	3	0.03
S. pyrenG	3	0.03
(S.torgal. x S. arad.)	6	0.06
S. torgal.	6	0.06
S. arad.	19	0.18
Others	41	0.38
total	107	1.00

The proportion of OGs under positive selection were lower than with Dataset C (~2%), which could be due to the shorter length of the alignments of this dataset in comparison to Dataset C (Figure 3.3). Of those 107 branches with signatures of selection, most were mapped to the ancestral of all species, as we found in Dataset C (41 pOGs, 38.3%). In general, most pOGs were under selection on the inner branches (the ancestral of our species) of the phylogeny, while the tip branches that correspond to selection acting on a single species had very few pOGs (Figure 3.9 and Tables 3.7 and 3.8). Probably because of this the *S. pyrenaicus* Guadiana and *S. torgalensis* did not show a higher percentage of pOGs, compared to northern

S. pyrenaicus Tagus and *S. carolitertii*, as inferred with Dataset C. Nevertheless, we detected a higher number of pOGs in *S. aradensis,* the southern species, with19 pOGs (17,8%) compared to less than 6 in the other species (Figure 3.9-II and Table 3.8). Contrary to Dataset C, both *S. pyrenaicus* populations had exactly the same number of pOGs (3), while *S. carolitertii* had slightly more (5 pOGs) (Figure 3.9-II and Table 3.8). However, we note that the species tree inferred for Dataset C was not exactly the same as the one inferred for Dataset D, since it had a polytomy on the clade of *S. carolitertii* and the two *S. pyrenaicus* and that the transcriptome assemblies used for the two *S. pyrenaicus* populations differed between the two datasets (see details in Material and Methods). For Dataset D we detected 15 pOGs in the ancestral population of *S. carolitertii* and both *S.pyrenaicus*, indicating that *S. carolitertii* and *S. pyrenaicus* share genes that were selected in the ancestral population.

3.5 - Functional enrichment analysis

3.5.1 Top score biological functions

To test if there were biological and molecular functions enriched on our set of pOGs, we inputted the *Danio rerio* annotations of the pOGs from Dataset C and D to DAVID. For Dataset C we found 23 clusters of functionally-related annotations enriched on the pOGs (Supplementary File 3.9). However, of these 23 functional clusters (FC), only the top scoring FC (related to the von Willerbrand factor type A) had a enrichment score above the 1.3 threshold (a threshold of 1.3 corresponds to a mean p-value of 0.05 for all the functionally related annotations that belong to a given cluster). This lack of significance was expected due to the low number of pOGs . Although we did not find more than one significant cluster, the annotation clusters are ranked by enrichment score and that still gives us information about the most important biological functions related to the pOGs we detected. Hence, we analysed the 10 top scoring functional clusters (Table 3.9). The top 10 clusters had annotations related with: von Willerbrand factor type A (5 pOGs); serine-type endopeptidase activity (7 pOGs); transferase of glycosyl groups (46 pOGs); zinc-finger, RING-type (25 pOGs); monooxygenase activity (77 pOGs), zinc-finger, C2H2 (40 pOGs), immune system process (45 pOG), integral component of the membrane (21 pOGs), muscle and neural development (45 pOGs) and apoptotic process (51 pOGs). For dataset D we obtained 5 functional clusters (Table 3.10 and Supplementary File 3.10), none with an enrichment score above the 1.3 threshold, likely due to the low number of pOGs. These clusters were: endopeptidase activity (28 pOGs), integral component of the membrane (21 pOGs), development and metabolic process regulation (57 pOGs), carboxylic acid metabolism (5 pOGs) and nucleic acid metabolism and nucleoside triphosphate binding (Table 3.10).

3.5.2. Biological functions under positive selection across the western Iberian *Squalius* species tree

By examining the ortholgous groups with significant positive selection (pOGs) belonging to each functional cluster, we could map the biological functions they represented to the species tree. Thus, for each functional cluster we obtained the branches of the phylogeny with pOGs that belonged to that cluster (Figure 3.10 and Tables 3.11 and 3.12). On most cases the proportion of pOGs from a given branch on a functional cluster seemed proportional to the total number of pOGs found on that given branch. To test if there was an interaction between the branch of the species tree and the functional cluster a given pOG belonged to, we performed a chi-square test and found no significant relationship (p-value = 0.86, Supplementary Figure 3.3). This indicates that there was no association between a specific function and a given branch of the species tree. Even so, there were some functional clusters where there was a clear deviation from the branch distribution of all the pOGs. This happened for the "von Willerbrand factor, type A" functional cluster, where 60% of the pOGs were under selection on the branch of the Guadiana *S. pyrenaicus*; and the "serine-type endopeptidase activity" where almost 60% of the pOGs were under selection on *S. torgalensis*. These deviations suggest that these functions were mainly selected on those specific species branches and therefore could be key factors for the adaptation of those species to their particular environment (Figure 3.10.I and Table 3.11). Interestingly, these are the two species of dataset C inhabiting the Mediterranean type climate.

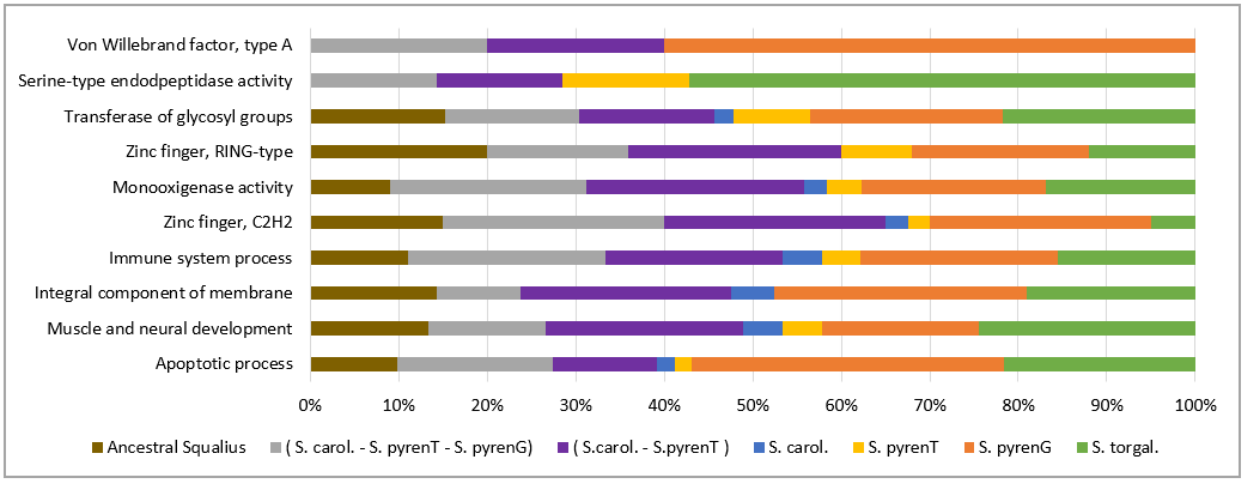
Table 3.9 - Top 10 scoring Functional Clusters inferred for Dataset C's pOG list as given by DAVID's Functional Clustering, ranked according to enrichment score.

Functional Cluster	Number of pOGs	Enrichment scores
Von Willebrand factor, type A	5	1.71
Serine-type endopeptidase activity	7	0.87
Transferase of glycosyl groups	46	0.84
Zinc finger, RING-type	25	0.76
Monooxigenase activity	77	0.72
Zinc finger, C2H2	40	0.71
Immune system process	45	0.54
Integral component of the membrane	21	0.46
Muscle and neural development	45	0.44
Apoptotic process	51	0.41

Table 3.10 – Top 5 Functional Clusters inferred for Dataset D's pOG list as given by DAVID's Functional Clustering Analysis, ranked according to enrichment score.

Functional Cluster	Number of pOGs	Enrichment scores
Endopeptidase activity	28	0.85
integral component of the membrane	21	0.39
development & metabolic process regulation	57	0.37
carboxylic acid metabolism	5	0.35
Nucleic acid metabolism /nucleoside triphosphate binding	25	0.34

I



II

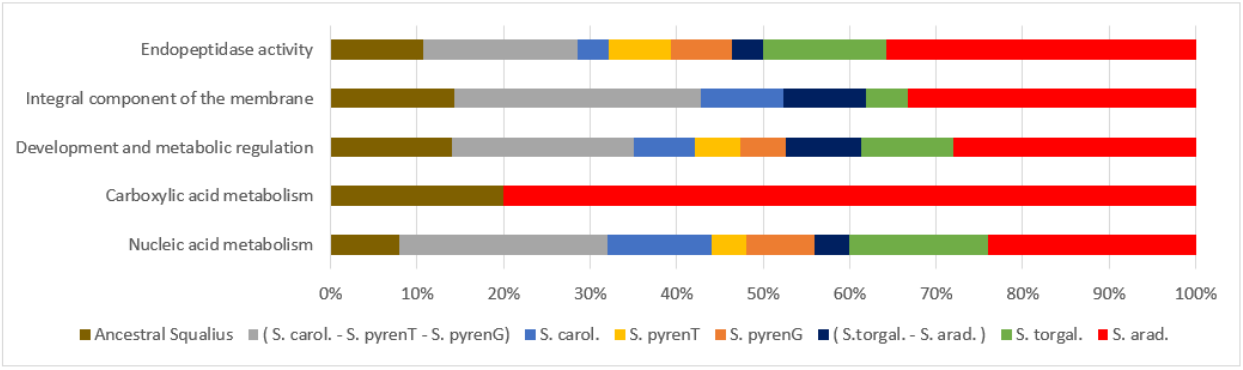


Figure 3.10 - Proportion of pOGs in each Functional Category that are under selection on each branch of the inferred species tree for I) Dataset C, II) Dataset D.

Table 3.11 - Number of pOGs in each Functional Category that are under selection on each branch of the species tree inferred for Dataset C. Species names within parentheses indicate the most recent common ancestor of the species within the parentheses.

Cluster	Ancestral Squalius	(S. carol. - S. pyrenT - S. pyrenG)	(S.carol. - S.pyrenT)	S. carol.	S. pyrenT	S. pyrenG	S. torgal.
Apoptotic process	5	9	6	1	1	18	11
Muscle and neural development	6	6	10	2	2	8	11
Integral component of membrane	3	2	5	1	0	6	4
Immune system process	5	10	9	2	2	10	7
Zinc finger, C2H2	6	10	10	1	1	10	2
Monooxigenase activity	7	17	19	2	3	16	13
Zinc finger, RING-type	5	4	6	0	2	5	3
Transferase of glycosyl groups	7	7	7	1	4	10	10
Serine-type endodpeptidase activity	0	1	1	0	1	0	4
Von Willebrand factor, type A	0	1	1	0	0	3	0

Table 3.12 - Number of pOGs in each Functional Category that are under selection on each branch of the species tree inferred for Dataset D. Species names within parentheses indicate the most recent common ancestor of the species within the parentheses.

Cluster	Ancestral Squalius	(S. carol. - S. pyrenT - S. pyrenG)	S. carol.	S. pyrenT	S. pyrenG	(S.torgal. - S. arad.)	S. torgal.	S. arad.
Nucleic acid metabolism	2	6	3	1	2	1	4	6
Carboxylic acid metabolism	1	0	0	0	0	0	0	4
Development and metabolic regulation	8	12	4	3	3	5	6	16
Integral component of the membrane	3	6	2	0	0	2	1	7
Endopeptidase activity	3	5	1	2	2	1	4	10

For Dataset D, we found a pattern similar to the one in Dataset C, where for most of the functional clusters, the proportion of pOGs from each branch did not deviate significantly from the proportion of total pOGs on the branches (Figure 3.10.II and Table 3.12, no significant relationship – chi-squared test p-value = 0.96). On Dataset D only the "carboxylic acid methabolism" cluster shows a clear deviation from the pattern of distribution of pOGs through the phylogeny. Interestingly, we found 80% of the pOGs related to this function in the *S. aradensis* branch, which is the species with a southernmost distribution. Thus, this could be related with the specific selection pressures of this species (Figure 3.10. II and Table 3.12).

3.6 – Target genes related to temperature response and circadian rhythm

Most of the bibliography genes involved in temperature response (Figures 2.1) were present both in Dataset C and D, however none of them showed signatures of positive selection under the normal aBSREL test using our datasets and final Pipeline 3 (Supplementary File 3.11). Interestingly, when running aBSREL without the Bonferroni correction we found signatures of selection in two genes: the *hsp90aa1* and *idh3b.* Even more remarkable, in both cases the signal of positive selection came from branches related, directly and indirectly, to *S. torgalensis* and *S. carolitertii,* the two species that were used on the studies where these genes were identified (Jesus et al., 2016, 2017). Despite being only significant when using uncorrected p-values, this result cannot be disregarded when considering the conservative nature of our pipeline. Regarding the genes related to the circadian rhythm, we were unable to find signatures of positive selection in any of them using our pipeline inclusively when removing the Bonferroni correction from aBSREL (Supplementary File 3.12). However, one important aspect is that Moreno, J., 2018 used sequences of these genes obtained by Sanger sequencing, which tend to have a better quality and to be more complete than the NGS sequences, especially in long genes like these. When comparing the sequences of these genes present in Dataset C (that had a greater mean alignment length than Dataset D) with the Sanger sequences used by Moreno, J., 2018, we found that our alignments were incomplete when compared with the ones used by Moreno, J., 2018 (Supplementary Figure 3.5 to 3.7).

4. Discussion

In this study we analysed transcriptomic data from several *Squalius* species distributed along different environments in Portugal. We tested several pipelines to process these data to detect orthologous genes, accounting for the possibility of different isoforms across species. Using a new conservative pipeline (Pipeline 3) that we developed to correct for alignment errors, we still found ~2% of genes with evidence of positive selection. This is close to the lower bound of the values in other studies using *RNA-seq* data and dN/dS methods, which vary greatly from 0% to 66% (Baker et al., 2018; Castoe et al., 2013; Cicconardi et al., 2017; Ghiselli et al., 2018; Lan et al., 2018; Tong et al., 2017; Xu et al., 2013; Yang et al., 2015) and are likely influenced by the organisms studied, the number of species used and the exact dN/dS test used. However, similar studies in bony fish found a proportion of genes with signatures of positive selection similar to what we found in our study, between 2% and 4% (Lan et al., 2018; Tong et al., 2017; Xu et al., 2013; Yang et al., 2015). Using a branch specific test, we were able to map such positively selected genes into the species tree and found signatures of selection at all points of the western Iberian *Squalius* clade. On the extant species we found support for a relatively higher number of genes under positive selection in southern populations and species, namely *S. pyrenaicus* from Guadiana, *S. torgalensis* and *S. aradensis*. A functional enrichment analysis suggests that there are several functional groups that could have been involved in adaptation in *Squalius* species, some specific to branches of the species tree. Interestingly, we found functional clusters with a higher proportion of pOGs with signatures of positive selection on southern species inhabiting the Mediterranean climate, namely Guadiana *S. pyrenaicus*, *S. torgalensis* and *S. arandensis*. Below we discuss these results. We start by discussing technical aspects regarding the new pipeline that we developed, and then we focus on the relevance of our results to understand the evolutionary history of *Squalius* species.

4.1 – New splicing-aware pipeline for comparative sequence analysis on transcriptomic data

4.1.1 - Development and improvement of the pipeline

The first and most challenging goal of this study was the development of a bioinformatic pipeline for a comparative genomics analyses using transcriptomic data from different studies and organs. This is different from common RNA-seq studies (Baker et al., 2018; Breschi et al., 2017; Ghiselli et al., 2018) because our goal was not to compare differences on gene expression between species but to obtain one coding sequence per gene for each species, and them compare them to test for signatures of positive selection along the species tree. Still, there are studies similar to ours that do comparative sequence analysis with RNA-seq (Baker et al., 2018; Cicconardi et al., 2017; Ghiselli et al., 2018; Wang et al.,

2017; Yang et al., 2012; Zhao et al., 2014). We will discuss how our bioinformatic pipelines compare with the pipelines used on two of these studies (Cicconardi et al., 2017; Wang et al., 2017). We compare our results with Wang et al. (2017) because their pipeline served as reference for the development of our first pipeline (Pipeline 1), and with Cicconardi et al. (2017) because their study is the most similar to ours. In Wang et al. (2017), their objective was to sequence the transcriptome of four luminescent beetles species using RNA extracted from whole body samples, and to use it for phylogenetic analysis. Cicconardi et al. (2017) studied the role of positive selection throughout the radiation of the genus *Drosophila* and looked for biological functions likely to be under positive selection in each lineage of the phylogeny. To our knowledge, Cicconardi et al. (2017) study is the closest to our methodology, using transcriptomic data from multiple species to test for signatures of selection across the coding transcriptome.

Using transcriptomic data for comparative sequence analysis poses a great challenge for the identification of orthologous sequences due to the ubiquitous presence of alternative splicing, which can bias the results of analysis if different splicing isoforms are accidentally analysed together (Zambelli et al., 2010). In our case this challenge was further increased by the fact that most of the transcriptomes we used were not sequenced specifically for this study (like in Wang et al. 2017), neither came from highly curated databases (like in Cicconardi et al. 2017). This meant that RNA was extracted from different organs under different conditions and therefore we could not expect the same splicing isoforms to be present on all the transcriptomes. Thus, our first goal was to develop a bioinformatic pipeline that could deal with this type of not-controlled transcriptomic data and output good quality alignments of orthologous sequences that could be used for further comparative genomics analysis. Our Pipeline 1 was based on the pipeline used by Wang et al. (2017) and is very similar to a conventional comparative genomics pipeline. We were able to recover a high number of orthologous groups (OGs) and we detected ~14% of OGs with signatures of positive selection (pOGs, Figure 3.2 and Table 3.5). This is higher than previously reported estimates for fish, with 2-4% of genes under positive selection (Lan et al., 2018; Tong et al., 2017; Xu et al., 2013; Yang et al., 2015). Furthermore, many genes had misaligned regions in one or two sequences (Figure 3.5 to 3.7), suggesting that many pOGs obtained in Pipeline 1 could be false positives due to poor alignment. Since these misaligned regions were restricted only to some regions of the alignment, we hypothesized that they most likely resulted from the presence of sequences from different transcript isoforms. The widely used method to remove misaligned regions GBLOCKS is not suited to detect such regions (see Material and Methods).

In Pipeline 2 we improved Pipeline 1 in two key steps: 1) to increase the accuracy of the ortholog identification step we used the OrthoDB suite, a specialized software package to find ortholog sequences between genomes/transcriptomes; 2) on the alignment cleaning step we complemented GBLOCKS with a custom script that removed the regions of the alignment where one or few of the sequences were completely misaligned from the rest (which would be putative alternative exons) outputting only

relatively conserved regions of the alignment. On Pipeline 2 we did not found any misalignments on the pOGs aligments, but on the other side we retrieved a very low number of OGs, indicating this pipeline is very conservative, i.e. it is reducing the changes of false positives at the cost of discarding true positives (Figure 3.2 and Table 3.5).

We suspected the low number of OGs retrieved by OrthoDB in Pipeline 2 could be related to two factors. First, we were very conservative by performing a redundancy removal step (with CD-HIT-EST) previous to the ortholog identification step with OrthoDB. This was included to filter all but one splicing isoform of each transcript, however CD-HIT-EST is also implemented on the OrthoDB pipeline, namely for the creation of clusters of orthologous sequences. Thus, it is possible that by filtering our transcriptomes with CD-HIT-EST previously to running OrthoDB we could be diminishing the power of OrthoDB to find orthologous groups. Second, we discarded all OGs with more than one sequence per species, thus keeping only those with exactly one sequence for all species. This approach is used in some studies to ensure single-copy orthologs (Cicconardi et al., 2017; Ghiselli et al., 2018). However, if CD-HIT-EST was not effective at removing splicing isoforms of a transcript, we would expect many single-copy OGs with several sequences per species due to splicing isoforms. Therefore, it is unlikely to retrieve OGs with only one sequence per species.

Our Pipeline 3 is similar to the one used in Ciccornardi et al. (2017), which is interesting since they also merged data from online repositories; they also used a specialized software for OG search; they also used GBLOCKS to clean the alignment and manually checked a set of random alignments to evaluate the effectiveness of the aligner and cleaning procedure; they used aBSREL to test positive selection at the branch level; and they also performed a functional enrichment analysis using DAVID. Cicconardi et al. (2017) also took in consideration the impact of alignments quality on the results and tested for this effect. However, they simply compared the proportion of terminal branches under selection with and without filtering with GBLOCKS. Although they found that GBLOCKS reduced the number of orthologs with signatures of selection, just as in Wang et al. (2017), they did not take in consideration the GBLOCKS limitation of missing regions with few misaligned sequences and the fact this is likely to happen in transcriptomic data due to different isoforms. Even though half of Cicconardi et al. (2017) transcriptomes come from the highly curated 12 Genome *Drosophila* Consortium, they also used transcriptomes from other sources that were less complete. Therefore, the possibility of OGs for which the same isoform were not present on all the 23 transcriptomes they used is not negligible, which in turn could lead to the formation of OGs with mixed isoforms and false positives genes under selection. Similarly, although Wang et al. (2017) generated RNA-seq data under the same conditions, hence expecting similar isoforms across their transcriptome assemblies, they used larvae individuals for the transcriptome of the ingroup species but an adult individual for the transcriptome of the outgroup. Developmental stage is one of the main factors of specificity of alternative splicing (Baralle and Giudice, 2017), therefore transcriptomes from different developmental stages are just as likely to include different

splicing isoforms like transcriptomes from different organs. Since Wang et al. (2017) only used GBLOCKS to clean their alignments, it is possible that some of the genes they detected to be under positive could be affected by misalignments, as was the case for our results from Pipeline 1.

Our results suggest that any study using transcriptomic data for comparative analysis should take in consideration the bias of mixing splicing isoforms within orthologous genes. We propose two ways to deal with this issue, either (i) by identifying removing them from the analysis or (ii) by using an approach of removing regions on the alignment with different exons, whether with the methodology we used in this study or a similar one. Our approach is aimed to studies that use transcriptomic data from public repositories and therefore not necessarily sequenced from the same organs nor in the same conditions. This type of data can have very different splicing patterns, and therefore taking the first approach could reduce severely the final number of OG in the dataset, making the second approach possibly a better alternative in those cases. With the increasing availability of data in public repositories, this type of meta-analyses combining different sources of transcriptomic data is expected to become more common in the near future. Further studies are, however, required to evaluate the performance of our proposed approach, namely to assess the power and accuracy to detect true positive genes under selection (see below).

4.1.3 – Current limitations to distinguish isoforms from paralogs

One of the remaining challenges with our Pipeline 3 is that it is still difficult to detect very similar paralogous genes. Some studies avoid this by working only with single-copy OGs (i.e. with one sequence per species) (Cicconardi et al., 2017; Ghiselli et al., 2018), as we did in Pipeline 2. However, results of Pipeline 2 indicate that such an approach was too conservative for our data (e.g. 475 OGs on Dataset B with Pipeline 2 vs 13,525 OGs and 9,605 on Dataset C and D, respectively with Pipeline 3, Table 3.5, Figure 3.2). However, keeping only single-copy OGs does not guarantee that our final dataset is paralog free. This is because gene duplication can lead to subfunctionalization of the gene copies (He and Zhang, 2005; Zhang, 2003), which can be either temporal (e.g. paralogs expressed on different developmental stages) or spatial (i.e. paralogs expressed on different tissues). If duplication are very recent, these are harder to identify than the alternative splicing isoforms, because we do not expect to see clusters of misaligned regions in specific regions of the alignment. Instead, depending on the time of divergence between the two paralogous and the strength of purifying selection, we expected gradual differences to accumulate between the paralogs of different species just like with true orthologs. This is very difficult to detect and avoid without good quality reference genomes or transcriptomes. Nevertheless, we only expect this to happen for very recent gene duplications. Given the lack of estimates about the proportion of duplicated genes and their age in the Western Iberian *Squalius* species,

at this stage we cannot be certain about the extent of this confounding factor. Therefore, we assumed that the results of OrthoDB are most likely due to alternative isoforms.

4.2 – Evolutionary history of Portuguese *Squalius* fish

4.2.1 - Dominant gene tree patterns across the transcriptome

Based on Dataset C, which lack the *S. aradensis* species but had longer alignments (Figure 3.3) we inferred the dominant gene tree cluster C1, with 39% of the OGs (Figure 3.4 and Table 3.6). This supported a closer relationship between *S. carolitertii* and the Tagus *S. pyrenaicus*, than between Tagus and Guadiana *S. pyrenaicus* populations, suggesting that *S. pyrenaicus* is paraphyletic in relation to *S. carolitertii*. Interestingly, this is in agreement with results of previous studies based on 7 nuclear genes (Sousa-Santos et al., 2019; Waap et al., 2011), suggesting that the patterns found on these nuclear markers are also found at the transcriptome level. The second most supported gene tree cluster (GTC C2 - 28%) had a median tree with a polytomy between all the *Squalius* species. This most likely results from lack of information on our alignments, which could be incomplete either because of our trimming process to remove alternative exons or because the sequences were already incomplete on some transcriptomes. Another alternative to explain the polytomy would be that a great proportion of genes are under very strong purifying selection across all the *Squalius* species.

The other gene tree clusters had very similar proportions, which is the expected pattern due to incomplete lineage sorting under neutrality (Nichols, 2001) (Supplementary Table 3.1, Supplementary Figure 3.1). Interestingly, GTC C5 clusters northern (*S. carolitertii* and the Tagus *S. pyrenaicus*) species in one clade and southern species (Guadiana *S. pyrenaicus* and *S. torgalensis*) in another clade. Given that these groups of species inhabit different climatic regimes (Atlantic in the north and Mediterranean in the south), likely imposing different selective pressures, an alternative explanation to incomplete lineage sorting is that this median topology reflects adaptive convergence of species in similar climate types. We do not find an overrepresentation of pOGs in gene tree cluster C5 (Supplementary Figure 3.4.I), but more complete alignment data would be required to confirm if convergence happened at least in some of the OGs with this topology.

For Dataset D, the most supported gene tree cluster (D1) represented 26.8% of the OGs and supported a gene tree consistent with the species tree but with a polytomy on the branch of *S. carolitertii* and the two *S. pyrenaicus* (Table 3.6, Figure 3.4-II). Surprisingly, in Dataset D we did not find any gene tree cluster supporting a clustering of the *Squalius* species by climate type, like in Dataset C. The differences between the gene tree of dataset C and D could be due to technical factors, such as the differences on the transcriptome assemblies used on both datasets (Table 2.2). The higher proportion of polytomies in Dataset D suggest that differences are very likely due to the higher quality of Dataset C, with almost twice the mean alignment length (600.7 bp on Dataset D vs 1014.5 bp on Dataset C). Smaller alignment

lengths on Dataset D were most likely due to the fragmented state of the *S. aradensis* transcriptome (Table 3.2 and Table 3.3). Overall, the gene trees we infer along the transcriptome are consistent with the previously reported species tree phylogeny, but also indicate that there is a high variation among genes. The fact that most of the incompatible gene trees topologies have similar proportions of OGs supporting them, suggest that they are due to ancestral polymorphism (incomplete lineage sorting) and not due to convergent selection.

4.2.2 – Signatures of positive selection on *Squalius* fishes

This is the first study to perform a transcriptome-wide scan for signatures of positive selection on the Portuguese *Squalius* species. We found signatures of positive selection in all branches of the species phylogeny, both in Dataset C and in Dataset D. Given that these are obligatory freshwater fish species, the speciation events on this clade were likely related with changes on the drainages of the Iberian Peninsula (Sousa-Santos et al., 2019). Since the estimated arrival of the ancestor of all western Iberian *Squalius* to the Iberian Peninsula 19 Mya ago (Sousa-Santos et al., 2019), the Iberian Peninsula basins suffered fragmentation events and changes between endorheic and exorheic regimes. These changes promoted the diversification of the Iberian *Squalius* clade by isolating populations, but it is also possible that they also changed the environment. Therefore, the presence of signatures of positive selection in all branches of the phylogeny could be evidence of a constant adaptation of the species of this clade to the changing conditions of the hydrographic network on Iberian Peninsula. Interestingly, on Dataset C we found comparable numbers of pOGs on all branches of the phylogeny with exception of *S. carolitertii* and the Tagus *S. pyrenaicus* (Table 3.7, Figure 3.9-I), which showed less number of genes. One hypothesis to explain this is that, since *S. carolitertii* and *S. pyrenaicus* diverged more recently than the other species (Sousa-Santos et al., 2019), they had less time to accumulate mutations for natural selection to act upon. However, an alternative hypothesis to explain these results is that the environment these species inhabit is similar to the environment of their ancestral. This is supported by the fact that the divergence of these two species is thought to have occurred when the Tagus and Douro basin became separated, and that the ancestral of these species is thought to have been present on both the Douro and the Tagus basins (Sousa-Santos et al., 2019). We can thus speculate that the formation of the Tagus and Douro basins did not change significantly the conditions on either drainage, and therefore the *S. carolitertii* and the Tagus *S. pyrenaicus* were not exposed to great selective pressures after their divergence.

Interestingly, in both Datasets C and D we found more genes under positive selection on the species under the Mediterranean climate, i.e. *S. torgalensis, S. pyrenaicus* Guadiana and *S. aradensis* (Figure 3.9-II, Table 3.8). Two recent studies comparing temperature response on *S. torgalensis* (a southern species under the Mediterranean climate type) and *S. carolitertii* (a northern species under the Atlantic climate type), showed that *S. torgalensis* gene expression was much less affected by changes in

temperature than in *S. carolitertii*, and also found that on *S. torgalensis* some proteins related to temperature response had changes that increased their thermostability, which led the authors to suggest that *S. torgalensis* could be adapted to the higher temperatures characteristic of its habitat (Jesus et al., 2016, 2017). The fact that we found signatures of selection (when using uncorrected p-values) in two of the genes identified in those studies, *hsp90aa1* and *idh3b* (Supplementary File 3.11) supports that hypothesis. However, despite these evidences of adaptation to temperature in *S. torgalensis*, it is likely that *S. aradensis* and the Guadiana population of *S. pyrenaicus* are exposed to even higher selective pressures due to higher temperatures. This is because the Mira river inhabited by *S. torgalensis* has substantial shading (Coelho, personal information), compared to the drainages inhabited by the other two species. Therefore, it is possible that the higher number of positive selection genes found on the Guadiana *S. pyrenaicus* and *S. aradensis* is associated with adaptation to high temperatures. Furthermore, the high number of pOGs in *S. aradensis* is especially significant because all the other species had very few pOGs for dataset D. Due to the shorter alignments of dataset D, we expect less power for dataset D. The fact that *S. aradensis* exhibited a relatively much higher number of pOGs, even with the limited dataset D suggest that the habitat of *S. aradensis* could entail stronger selective pressures than the habitat of the other *Squalius* species in this study. Obtainaing a high quality transcriptome for *S. aradensis* would be required to further elucidate if this species has evidence for stronger selective pressures, involving a response in more genes and pathways than in the other species inhabiting the Mediterranean climate type.

In total, we inferred between 1.4% or 2.0% of the genes with signatures of positive selection (datasets D and C, respectively). This is comparable with values reported in similar studies with fish species, where estimates point to 2% to 4% of ortholog genes with signatures of positive selection (Lan et al., 2018; Tong et al., 2017; Xu et al., 2013; Yang et al., 2015). However, it is important to note that due to the relatively short length of our resulting alignments and the conservative nature of Pipeline 3, it is likely that we are underestimating the actual proportion of genes under positive selection. This was clearly shown when we compared our results with the ones obtained in a previous study that used high quality Sanger sequencing of circadian genes (Moreno, J., 2018). We did not find any signatures of selection in our transcriptome alignments for circadian genes that were under positive selection with the Sanger data. Comparing our alignments with the ones obtained by Moreno, J. (2018), we found that our alignments were incomplete (Suplementary Figures 3.5-3.7), either because some sequences were incomplete on the transcriptome or because a region of the alignment had been removed by our cleaning process. This showed that at least in some cases we missed true positives due to the incompleteness of the sequences on the transcriptome or due to the mixture of different splicing isoforms on the alignments.

Another important aspect to note is that the highest number of genes under selection was found at the base of the phylogeny, corresponding to the ancestral of all *Squalius* species and one of the outgroup species, either *D. rerio* or *L. burdigalensis*. The signal in this point of the phylogeny is mostly due to

pOGs with gene trees clustering one of the outgroup species to one of the ingroup species (Suplementary Figure 3.1). The reasons for this type of positive selection in the ancestral branch could be due to: 1) adaptations on the ancestral of all our species followed by incomplete lineage sorting of those beneficial alleles in different species; 2) adaptative convergence between the ingroup species and the outgroup species; or 3) false positives related with the sequence of those species having a splicing isoform different from the rest of the species on the alignment. Even though our pipeline 3 should remove the effects of the third point, more transcriptomic data from *Leusciscinae* species, such that we would cover the gap between the ingroup and outgroup, would be required to clarify among these alternative hypotheses.

4.2.3 - Biological functions under selection throughout the evolution of the Iberian *Squalius* fishes

We were able to find 80 functional clusters of annotations in Dataset C and 5 in Dataset D. We were only able to find one significantly enriched functional cluster in Dataset C – related to the van Willerbrand factor, type A – and none of the functional clusters in Dataset D reached statistical significance. This difference between the two datasets is probably related to the lower number of pOGs present on Dataset D, which limited the power of the functional analysis. We were unable to find any branch significantly enriched for a particular function (data not shown), probably because of the low number of significant positive selected pOGs we inferred for each branch (much less than 100 the minimum number of genes suggested by authors of DAVID (Huang et al., 2009).

The fact that even with an approach that maximized the power of the enrichment analysis we only found one enriched cluster on Dataset C and none in Dataset D probably means either that 1) there are almost no biological functions that have been consistently under selection throughout the evolution of these species or 2) if there are, the signal is too weak in our datasets due to the conservative filtering and trimming of our pipeline that resulted in a small set of pOGs available for the analysis.

Still, we found some functional clusters that can give insigths about adaptation in the southern species that inhabit the Mediterranean climate. For Dataset C the functional cluster related to the von Willerbrand factor, type A was significantly enriched. This function is also disproportionally found to be associated with the branch of southern *S. pyrenaicus* Guadiana (Figure 3.10 and Tables 3.11). This factor is a multimeric glycoprotein found in blood plasma and is involved on platelet adhesion and haemorrhage coagulation. In humans mutations on the gene coding this factor are related with bleeding disorders (Bharati and Prashanth, 2011; Ruggeri and Ware, 1993). Interestingly, 3 out of 5 of the positive selected genes found on this functional cluster were under selection on the Guadiana *S. pyrenaicus*, whose habitat is under a Mediterranean climate-type, which is characterized by high temperatures and droughts in summer. Blood viscosity varies inversely with temperature in humans and fishes (Çinar,

2001; Eckmann et al., 2000; Graham and Fletcher, 1983; Graham et al., 1985; Rand et al., 1964; Wells et al., 1990), thus it is possible that the high summer temperatures could be lowering the blood viscosity of the Guadiana *S. pyrenaicus* populations, which could increase haemorrhage closure times. Modifications on clotting factors like the von Willerbrand factor A could thus be favoured by natural selection. However, we do not find any term related to the von Willerbrand factor on *S. aradensis*, whose habitat is similar to the one of the Guadiana *S. pyrenaicus*. As mentioned above, the Dataset D with *S. aradensis* had incomplete alignments and a low number of pOGs, and hence we cannot be sure whether the absence of this functional cluster indicates that this species responded differently when faced with similar selective pressures or if we just did not have enough power on Dataset D to detect this signal on *S. aradensis* .

There were two other functional clusters that, despite not being signicantly enriched, had a disproportionate number of pOGs on particular branches of the species tree. One of this clusters was the "serine-type endopeptidase activity" cluster which had almost 60% of the pOGs on *S. torgalensis* (Figure 3.10.I and Table 3.11). Serine-type endopeptidases are a big family of peptidases that participate in a very wide array of functions, including homeostasis, immune response, blood coagulation and signalling (Hedstrom, 2002). On Dataset D the "carboxylic acid methabolism" cluster shows a clear deviation from the pattern of distribution of pOGs through the phylogeny, with 80% of the pOGs related to this function in the *S. aradensis* branch (Figure 3.10. II and Table 3.12). These results could be related with to specific selection pressures on *S. torgalensis* and *S. aradensis*, however due to the generic nature of both functional clusters, it is difficult to speculate about the selective pressures related to these genes. Since both functional clusters have few pOGs (Table 3.9 and Table 3.10) it is also possible that these deviations from the pattern of distribution of the pOGs are artefacts due to the low sample size.

Though not significantly enriched, another interesting functional cluster was the one related with immune system process, which was found in similar proportions in all branches of the species tree. Recent studies have found that the gene expression of some immune genes is much more robust to changes on temperature and acidification on *S. torgalensis* than in *S. carolitertii*, and inclusively that there are structural differences on one protein (GBP1) between the two species that give an increased thermostability to the protein on *S. torgalensis*, suggesting a possible adaptation to temperature (Jesus et al., 2016, 2017). The fact that we found a functional cluster related to immune system processes reinforces the idea that response to increasing temperatures in these fish species might involve genes of the immune system. Our results suggest that those immunity genes are not restricted only to *S. torgalensis*. Further studies would be required to analyse in more detail immunity genes in each species.

The presence of functional clusters related to development ("muscle and neural development", "development" and "apoptosis") at first sight might seems intriguing since the transcriptomes we used were mostly from adult individuals. However, developmental genes like the Hox genes can have

functions beyond the developmental phase of organisms (Gavin et al., 1990; Rux and Wellik, 2017) so it is not completely surprising to find them expressed on adult individuals. Furthermore, since the western Iberian *Squalius* present a cline of external morphology traits, namely size, number lateral line scales, number of fin rays and number of gill rakers (Coelho et al., 1998), it is not surprising that genes related to development show signatures of selection. However, once again, further studies are needed to analyse the exact genes that belong to each functional group to identify which parts of the development were under selection on each species.

The rest of the functional clusters that were enriched are equally generic, and mostly related to proteolytic activity ("serine-type endopeptidase activity", "monooxygenase activity", "endopeptidase activity"), binding between molecules ("zinc-finger, RING-type" and "zinc-finger, C2H2"), metabolism ("development & metabolic process regulation","nucleic acid metabolism" ,"development and metabolic process regulation", "carboxylic acid methabolism ") and membrane components ("integral component of the membrane" found on both Dataset C and D). In most cases, the distribution of the pOGs in a functional cluster across the species tree matches the distribution of the set of all the pOGs, i.e the functional clusters show the same proportion of pOGs coming from each branch of the species tree as the total proportion of pOGs. In fact, contrary to what we expected, there was no evidence for a significant relation between the branch of the phylogeny and a particular set of functional clusters (chi-square test, Suplementary Figure 3.3). This means that, in general, the biological functions represented by our functional clusters are all following the same pattern of relative strength of the selection across the phylogeny. If this is a significant biological result it indicates that there were no unique biological functions under selection in any branch of the species tree, i.e. natural selection acted on the same biological functions across the species tree. However, given that there were not enough pOGs on each branch, we have limited power to discriminate signatures of putative unique biological functions selected specifically for a given branch of the phylogeny. Transcriptomic data from other *Squalius* species or more populations of *Squalius* would increase the power of our analyses and help clarify these aspects. The study of the specific genes represented by the pOGs on each branch would also allow to know more on where natural selection acted on each species.

5. Final remarks and Future Perspectives

In this study, we analysed for the first time the patterns of positive selection across the transcriptome of the western Iberian *Squalius* fish. We found more orthologs under positive selection on the branches of the extant species under the Mediterranean climate type (*S. aradensis, S. torgalensis* and the Guadiana *S. pyrenaicus*) than on the northern species under the Atlantic climate type (*S. carolitertii* and Tagus *S. pyrenaicus*). This suggests that species under the southern Mediterranean climate type were under stronger selective pressures. We also identified biological functions that have been targeted by natural

selection on this clade, which included immunity, development, proteolysis and blood coagulation, among others. Another aspect we looked into was the relationships between the species across the transcriptome, finding support for the paraphyly of *S. pyrenaicus* in relation to *S. carolitertii*, in agreement with recent phylogenetic studies using seven nuclear markers. We also developed a new bioinformatic pipeline for comparative analysis on transcriptome data that identifies alignments likely with different splicing isoforms and trims the regions corresponding to different exons, instead of removing the whole alignment, thus minimizing the loss of information due to this issue. Thus, our study provides new data and bioinformatic tools that can be used for future studies, namely two new transcriptomes for the western Iberian *Squalius* (*S. aradensis* and the Tagus *S. pyrenaicus*), hundreds of target genes with signatures of positive selection (potentially involved on the adaptation of these species), two datasets with thousands of putative orthologs between these species and their corresponding gene trees, and a bioinformatic pipeline with a new approach for comparative analysis on transcriptomic data. Finally, given that *Squalius* species cover two climatic types (Atlantic and Mediterranean), which have different temperatures and precipitation patterns, the information found in this study about genes under positive selection can help predict their potential to adapt to future environmental changes and have implications for management and conservation of these endangered species.

Future studies should focus on increasing the power of the analysis by obtaining more transcriptomes, specially a good quality transcriptome for *S. aradensis* but also transcriptomes from other species, like for example eastern Iberian *Squalius*, which were not included on this study. A reference genome from a close species could also be a very valuable asset to improve quality of the transcriptome assemblies and the identification of paralog genes and splicing isoforms. Indeed, all the transcriptome assemblies in this study were obtained *de novo* without a reference genome. A reference genome from a phylogenetically close species, equally distance from all *Squalius* could help obtaining assemblies maximizing the number of orthologous regions. Populational data would also be very valuable because it would allow to perform a much wider array of selection tests besides tests based on dN/dS, which could also detect selection on non-coding regions. An important selective force that was not investigated here was purifying selection. The dN/dS based tests can be used to detect purifying selection, however this was not implemented in aBSREL. In the future it would be interesting to characterize the patterns of purifying selection through the phylogeny of these species and see how they compare with the patterns of positive selection. This would allow us to gain insights both of the time and mode of how natural selection acted on the phylogeny of the western Iberian *Squalius.*

Another important information to understand the molecular basis of adaptation is detailed phenotypic and environmental data for each species, in order to truly be able to link the patterns of positive selection we find at the transcriptome level to adaptation of these species to their environment.

6. Bibliography

Baker, E.A.G., Wegrzyn, J.L., Sezen, U.U., Falk, T., Maloney, P.E., Vogler, D.R., Delfino-Mix, A., Jensen, C., Mitton, J., Wright, J., et al. (2018). Comparative Transcriptomics Among Four White Pine Species. G3amp58 GenesGenomesGenetics *8*, 1461–1474.

Baralle, F.E., and Giudice, J. (2017). Alternative splicing as a regulator of development and tissue identity. Nat. Rev. Mol. Cell Biol. *18*, 437–451.

Barrett, R.D.H., and Hoekstra, H.E. (2011). Molecular spandrels: tests of adaptation at the genetic level. Nat. Rev. Genet. *12*, 767–780.

Bharati, K.P., and Prashanth, U.R. (2011). Von Willebrand disease: an overview. Indian J. Pharm. Sci. *73*, 7–16.

Breschi, A., Gingeras, T.R., and Guigó, R. (2017). Comparative transcriptomics in human and mouse. Nat. Rev. Genet. *18*, 425–440.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421.

Carvalho, S.B., Brito, J.C., Crespo, E.J., and Possingham, H.P. (2010). From climate change predictions to actions - conserving vulnerable animal groups in hotspots at a regional scale: FROM CLIMATE CHANGE PREDICTIONS TO ACTIONS. Glob. Change Biol. *16*, 3257–3270.

Castoe, T.A., de Koning, A.P.J., Hall, K.T., Card, D.C., Schield, D.R., Fujita, M.K., Ruggiero, R.P., Degner, J.F., Daza, J.M., Gu, W., et al. (2013). The Burmese python genome reveals the molecular basis for extreme adaptation in snakes. Proc. Natl. Acad. Sci. *110*, 20645–20650.

Castresana, J. (2000). Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. Mol. Biol. Evol. *17*, 540–552.

Cicconardi, F., Marcatili, P., Arthofer, W., Schlick-Steiner, B.C., and Steiner, F.M. (2017). Positive diversifying selection is a pervasive adaptive force throughout the Drosophila radiation. Mol. Phylogenet. Evol. *112*, 230–243.

Çinar, Y. (2001). Blood viscosity and blood pressure: role of temperature and hyperglycemia. Am. J. Hypertens. *14*, 433–438.

Coelho, M.M., Brito, R.M., Pacheco, T.R., Figueiredo, D., and Pires, A.M. (1995). Genetic variation and divergence of Leuciscus pyrenaicus and L. carolitertii (Pisces, Cyprinidae). J. Fish Biol. *47*, 243–258.

Coelho, M.M., Bogutskaya, N.G., Aodrigues, J.A., and Collares-Pereira, M.J. (1998). Leuciscus torgalensis, and L. aradensis, two new cyprinids for Portuguese fresh waters. J. Fish Biol. *52*, 937–950.

Daub, J.T., Hofer, T., Cutivet, E., Dupanloup, I., Quintana-Murci, L., Robinson-Rechavi, M., and Excoffier, L. (2013). Evidence for Polygenic Adaptation to Pathogens in the Human Genome. Mol. Biol. Evol. *30*, 1544–1558.

Eckmann, D.M., Bowers, S., Stecker, M., and Cheung, A.T. (2000). Hematocrit, Volume Expander, Temperature, and Shear Rate Effects on Blood Viscosity: Anesth. Analg. *91*, 539–545.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797.

Foll, M., Gaggiotti, O.E., Daub, J.T., Vatsiou, A., and Excoffier, L. (2014). Widespread Signals of Convergent Adaptation to High Altitude in Asia and America. Am. J. Hum. Genet. *95*, 394–407.

Fresno, C., and Fernandez, E.A. (2013). RDAVIDWebService: a versatile R interface to DAVID. Bioinformatics *29*, 2810–2811.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150–3152.

Gavin, B.J., McMahon, J.A., and McMahon, A.P. (1990). Expression of multiple novel Wnt-1/int-1-related genes during fetal and adult mouse development. Genes Dev. *4*, 2319–2332.

Genomic Resources Development Consortium, Almeida-Val, V.M.F., Boscari, E., Coelho, M.M., Congiu, L., Grapputo, A., Grosso, A.R., Jesus, T.F., Luebert, F., Mansion, G., et al. (2015a). Genomic Resources Notes accepted 1 April 2015 - 31 May 2015. Mol. Ecol. Resour. *15*, 1256–1257.

Genomic Resources Development Consortium, Álvarez, P., Arthofer, W., Coelho, M.M., Conklin, D., Estonba, A., Grosso, A.R., Helyar, S.J., Langa, J., Machado, M.P., et al. (2015b). Genomic Resources Notes Accepted 1 June 2015 - 31 July 2015. Mol. Ecol. Resour. *15*, 1510–1512.

Genomic Resources Development Consortium, Blanchet, S., Bouchez, O., Chapman, C.A., Etter, P.D., Goldberg, T.L., Johnson, E.A., Jones, J.H., Loot, G., Omeja, P.A., et al. (2015c). Genomic resources notes accepted 1 December 2014 - 31 January 2015. Mol. Ecol. Resour. *15*, 684.

Ghiselli, F., Iannello, M., Puccio, G., Chang, P.L., Plazzi, F., Nuzhdin, S.V., and Passamonti, M. (2018). Comparative Transcriptomics in Two Bivalve Species Offers Different Perspectives on the Evolution of Sex-Biased Genes. Genome Biol. Evol. *10*, 1389–1402.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. *29*, 644–652.

Graham, M.S., and Fletcher, G.L. (1983). Blood and plasma viscosity of winter flounder: influence of temperature, red cell concentration, and shear rate. Can. J. Zool. *61*, 2344–2350.

Graham, M.S., Fletcher, G.L., and Haedrich, R.L. (1985). Blood viscosity in arctic fishes. J. Exp. Zool. *234*, 157–160.

He, X., and Zhang, J. (2005). Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics *169*, 1157–1164.

Hedstrom, L. (2002). Serine Protease Mechanism and Specificity. Chem. Rev. *102*, 4501–4524.

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

Howe, K., Clark, M.D., Torroja, C.F., Torrance, J., Berthelot, C., Muffato, M., Collins, J.E., Humphray, S., McLaren, K., Matthews, L., et al. (2013). The zebrafish reference genome sequence and its relationship to the human genome. Nature *496*, 498–503.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. *4*, 44–57.

Jeffares, D.C., Tomiczek, B., Sojo, V., and dos Reis, M. (2015). A Beginners Guide to Estimating the Non-synonymous to Synonymous Rate Ratio of all Protein-Coding Genes in a Genome. In Parasite Genomics Protocols, C. Peacock, ed. (New York, NY: Springer New York), pp. 65–90.

Jesus, T.F., Grosso, A.R., Almeida-Val, V.M.F., and Coelho, M.M. (2016). Transcriptome profiling of two Iberian freshwater fish exposed to thermal stress. J. Therm. Biol. *55*, 54–61.

Jesus, T.F., Moreno, J.M., Repolho, T., Athanasiadis, A., Rosa, R., Almeida-Val, V.M.F., and Coelho, M.M. (2017). Protein analysis and gene expression indicate differential vulnerability of Iberian fish species under a climate change scenario. PLOS ONE *12*, e0181325.

Jombart, T., Kendall, M., Almagro-Garcia, J., and Colijn, C. (2017). Treespace: Statistical exploration of landscapes of phylogenetic trees. Mol. Ecol. Resour. *17*, 1385–1392.

Jukes, T.H., and Cantor, C.R. (1969). Evolution of Protein Molecules. In Mammalian Protein Metabolism, (Elsevier), pp. 21–132.

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. *30*, 3059–3066.

Kong, F., Li, H., Sun, P., Zhou, Y., and Mao, Y. (2014). De novo assembly and characterization of the transcriptome of seagrass Zostera marina using illumina paired-end sequencing. PLoS ONE *9*, 1–19.

Kriventseva, E.V., Tegenfeldt, F., Petty, T.J., Waterhouse, R.M., Simão, F.A., Pozdnyakov, I.A., Ioannidis, P., and Zdobnov, E.M. (2015). OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. Nucleic Acids Res. *43*, D250–D256.

Lan, Y., Sun, J., Xu, T., Chen, C., Tian, R., Qiu, J.-W., and Qian, P.-Y. (2018). De novo transcriptome assembly and positive selection analysis of an individual deep-sea fish. BMC Genomics *19*.

Lassmann, T., and Sonnhammer, E.L.L. (2005). Kalign--an accurate and fast multiple sequence alignment algorithm. BMC Bioinformatics *6*, 298.

Leinonen, R., Sugawara, H., Shumway, M., and International Nucleotide Sequence Database Collaboration (2011). The sequence read archive. Nucleic Acids Res. *39*, D19-21.

Machado, M., Pinho, J., Grosso, A.R., Schartl, M., and Coelho, M.M. (2015). De novo assembled transcriptome of organs involved in reproduction in an endangered endemic Iberian cyprinid fish (Squalius pyrenaicus). Mol. Ecol. Resour. *15*, 1510–1512.

Magalhaes, M.F., Schlosser, I.J., and Collares-Pereira, M.J. (2003). The role of life history in the relationship between population dynamics and environmental variability in two Mediterranean stream fishes. J. Fish Biol. *63*, 300–317.

Matos, I., Machado, M.P., Schartl, M., and Coelho, M.M. (2015). Gene Expression Dosage Regulation in an Allopolyploid Fish. PLOS ONE *10*, e0116309.

Mesquita, N., Hanfling, B., Carvalho, G., and Coelho, M.M. (2005). Phylogeography of the cyprinid Squalius aradensis and implications for conservation of the endemic freshwater fauna of southern Portugal. 1939–1954.

Nichols, R. (2001). Gene trees and species trees are not the same. Trends Ecol. Evol. *16*, 358–364.

Nielsen, R. (2005). Molecular Signatures of Natural Selection. Annu. Rev. Genet. *39*, 197–218.

Notredame, C., Higgins, D.G., and Heringa, J. (2000). T-coffee: a novel method for fast and accurate multiple sequence alignment 1 1Edited by J. Thornton. J. Mol. Biol. *302*, 205–217.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet. *40*, 1413–1415.

Pond, S.L.K., Frost, S.D.W., and Muse, S.V. (2005). HyPhy: hypothesis testing using phylogenies. Bioinformatics *21*, 676–679.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLoS ONE *5*, e9490.

Rand, P.W., Lacombe, E., Hunt, H.E., and Austin, W.H. (1964). Viscosity of normal human blood under normothermic and hypothermic conditions. J. Appl. Physiol. *19*, 117–122.

Rosenberg, N.A., and Nordborg, M. (2002). Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nat. Rev. Genet. *3*, 380–390.

Ruggeri, Z.M., and Ware, J. (1993). von Willebrand factor. FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol. *7*, 308–316.

Rux, D.R., and Wellik, D.M. (2017). *Hox* genes in the adult skeleton: Novel functions beyond embryonic development: *Hox* Genes in the Adult Skeleton. Dev. Dyn. *246*, 310–317.

Savolainen, O., Lascoux, M., and Merilä, J. (2013). Ecological genomics of local adaptation. Nat. Rev. Genet. *14*, 807–820.

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

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinforma. Oxf. Engl. *31*, 3210–3212.

Smith, M.D., Wertheim, J.O., Weaver, S., Murrell, B., Scheffler, K., and Kosakovsky Pond, S.L. (2015). Less Is More: An Adaptive Branch-Site Random Effects Model for Efficient Detection of Episodic Diversifying Selection. Mol. Biol. Evol. *32*, 1342–1353.

Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J.M., and Kelly, S. (2016). TransRate: reference-free quality assessment of de novo transcriptome assemblies. Genome Res. *26*, 1134–1144.

Sousa-Santos, C., Jesus, T.F., Fernandes, C., Robalo, J.I., and Coelho, M.M. (2019). Fish diversification at the pace of geomorphological changes: evolutionary history of western Iberian Leuciscinae

(Teleostei: Leuciscidae) inferred from multilocus sequence data. Mol. Phylogenet. Evol. *133*, 263–285.

Stapley, J., Reger, J., Feulner, P.G.D., Smadja, C., Galindo, J., Ekblom, R., Bennison, C., Ball, A.D., Beckerman, A.P., and Slate, J. (2010). Adaptation genomics: the next generation. Trends Ecol. Evol. *25*, 705–712.

Stout, C.C., Tan, M., Lemmon, A.R., Lemmon, E.M., and Armbruster, J.W. (2016). Resolving Cypriniformes relationships using an anchored enrichment approach. BMC Evol. Biol. *16*, 1–13.

Talavera, G., and Castresana, J. (2007). Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments. Syst. Biol. *56*, 564–577.

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

Tong, C., Tian, F., and Zhao, K. (2017). Genomic signature of highland adaptation in fish: a case study in Tibetan Schizothoracinae species. BMC Genomics *18*.

Waap, S., Amaral, A.R., Gomes, B., Coelho, M.M., and Á, B.Á.M. (2011). Multi-locus species tree of the chub genus Squalius (Leuciscinae : Cyprinidae) from western Iberia : new insights into its evolutionary history.

Wallace, I.M. (2006). M-Coffee: combining multiple sequence alignment methods with T-Coffee. Nucleic Acids Res. *34*, 1692–1699.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470–476.

Wang, K., Hong, W., Jiao, H., and Zhao, H. (2017). Transcriptome sequencing and phylogenetic analysis of four species of luminescent beetles. Sci. Rep. *7*, 1–11.

Waterhouse, R.M., Seppey, M., Simão, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E.V., and Zdobnov, E.M. (2017). BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol. Biol. Evol.

Wells, R.M.G., Macdonald, J.A., and diPrisco, G. (1990). Thin-blooded Antarctic fishes: a rheological comparison of the haemoglobin-free icefishes Chionodraco kathleenae and Cryodraco antarcticus with a red-blooded nototheniid, Pagothenia bernacchii. J. Fish Biol. *36*, 595–609.

Xu, J., Ji, P., Wang, B., Zhao, L., Wang, J., Zhao, Z., Zhang, Y., Li, J., Xu, P., and Sun, X. (2013). Transcriptome Sequencing and Analysis of Wild Amur Ide (Leuciscus waleckii) Inhabiting an Extreme Alkaline-Saline Lake Reveals Insights into Stress Adaptation. PLoS ONE *8*, 1–10.

Yang, L., Wang, Y., Zhang, Z., and He, S. (2015). Comprehensive Transcriptome Analysis Reveals Accelerated Genic Evolution in a Tibet Fish, Gymnodiptychus pachycheilus. Genome Biol. Evol. *7*, 251–261.

Yang, W., Qi, Y., Bi, K., and Fu, J. (2012). Toward understanding the genetic basis of adaptation to high-elevation life in poikilothermic species: a comparative transcriptomic analysis of two ranid frogs, Rana chensinensis and R. kukunoris. BMC Genomics *13*, 588.

Zambelli, F., Pavesi, G., Gissi, C., Horner, D.S., and Pesole, G. (2010). Assessment of orthologous splicing isoforms in human and mouse orthologous genes. BMC Genomics *11*, 534.

Zdobnov, E.M., Tegenfeldt, F., Kuznetsov, D., Waterhouse, R.M., Simão, F.A., Ioannidis, P., Seppey, M., Loetscher, A., and Kriventseva, E.V. (2017). OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. Nucleic Acids Res. *45*, D744–D749.

Zhang, J. (2003). Evolution by gene duplication: an update. Trends Ecol. Evol. *18*, 292–298.

Zhao, X., Yu, H., Kong, L., Liu, S., and Li, Q. (2014). Comparative transcriptome analysis of two oysters, Crassostrea gigas and Crassostrea hongkongensis provides insights into adaptation to hypo-osmotic conditions. PloS One *9*, e111915.

7. Supplementary Material

Supplementary Table 2.2 – Target genes related to temperature response according to previous studies. Functional Category is the main biological process where each gene is involved.

Genes	Study	Differences	Functional Category
fkbp4	Jesus et al., 2016 & Jesus et al.,2017	gene expression & protein structure	protein folding
hsp90aa1.1	Jesus et al., 2016 & Jesus et al.,2017	gene expression and protein thermostability	protein folding
hsp90aa1.2	Jesus et al., 2016	gene expression and protein thermostability	protein folding
hsp70	Jesus et al., 2016	gene expression	protein folding
hsc70	Jesus et al., 2016 & Jesus et al.,2017	gene expression & protein structure	protein folding
uri1	Jesus et al., 2016	gene expression	protein folding
stip1	Jesus et al., 2017	gene expression	protein folding
gpx4b	Jesus et al., 2016	gene expression	oxidation-reduction process
hsd17b7	Jesus et al., 2016	gene expression	oxidation-reduction process
idh1	Jesus et al., 2016	gene expression	oxidation-reduction process
idh3b	Jesus et al., 2016	gene expression	oxidation-reduction process
ldha	Jesus et al., 2016	gene expression	oxidation-reduction process
ndufb8	Jesus et al., 2016	gene expression	oxidation-reduction process
glula	Jesus et al., 2016	gene expression	gluthamine biosynthesis
glulb	Jesus et al., 2016	gene expression	gluthamine biosynthesis
snrpd2	Jesus et al., 2016	gene expression	nucleic acid methabolism
pparab	Jesus et al., 2016	gene expression	regulation of transcription
gbp1	Jesus et al., 2016 & Jesus et al.,2017	gene expression, protein structure and thermostability	immune response

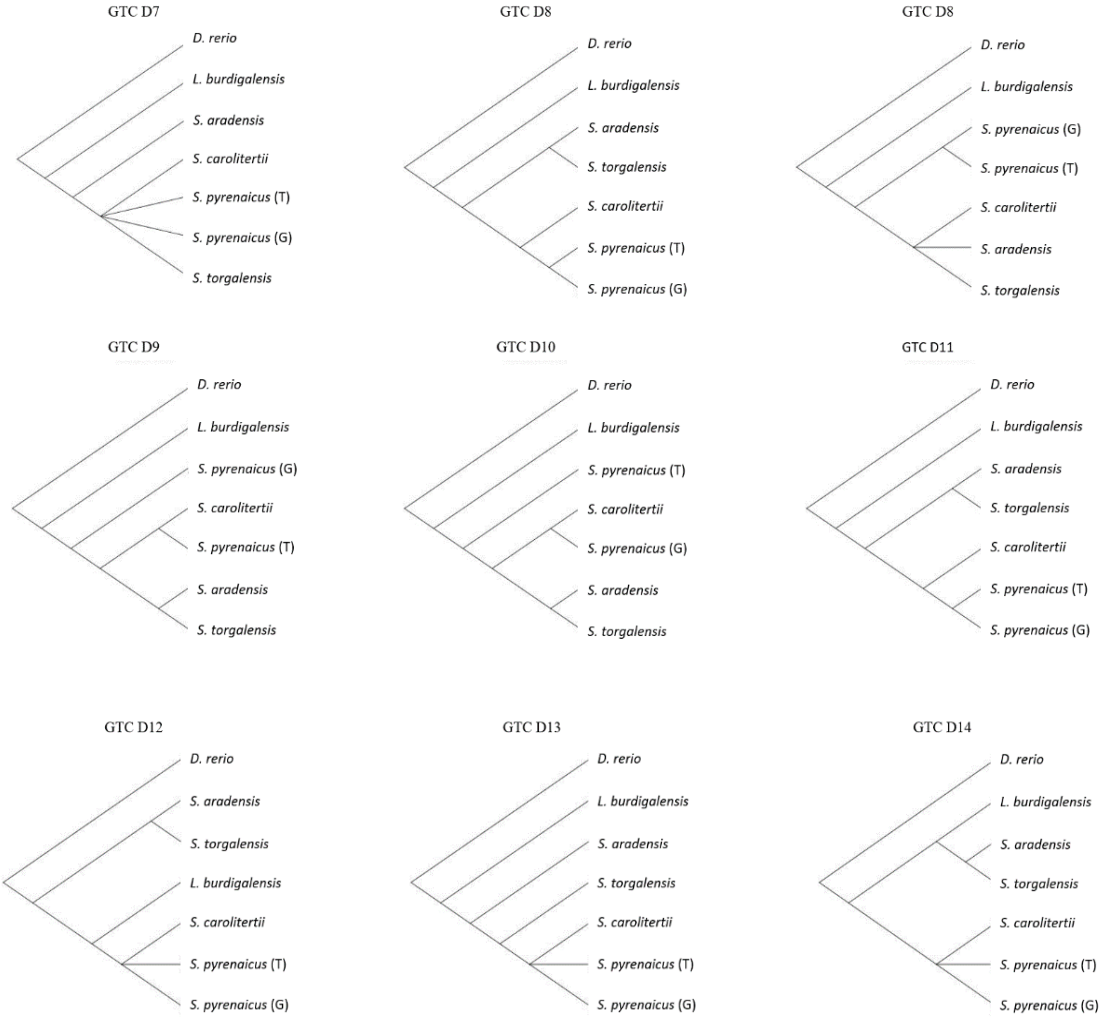
segf

Supplementary Table 2.2 – Circadian rhythm genes where signatures of positive selection using a branch-site dN/dS test where found on a previous study (Moreno, J., 2018).

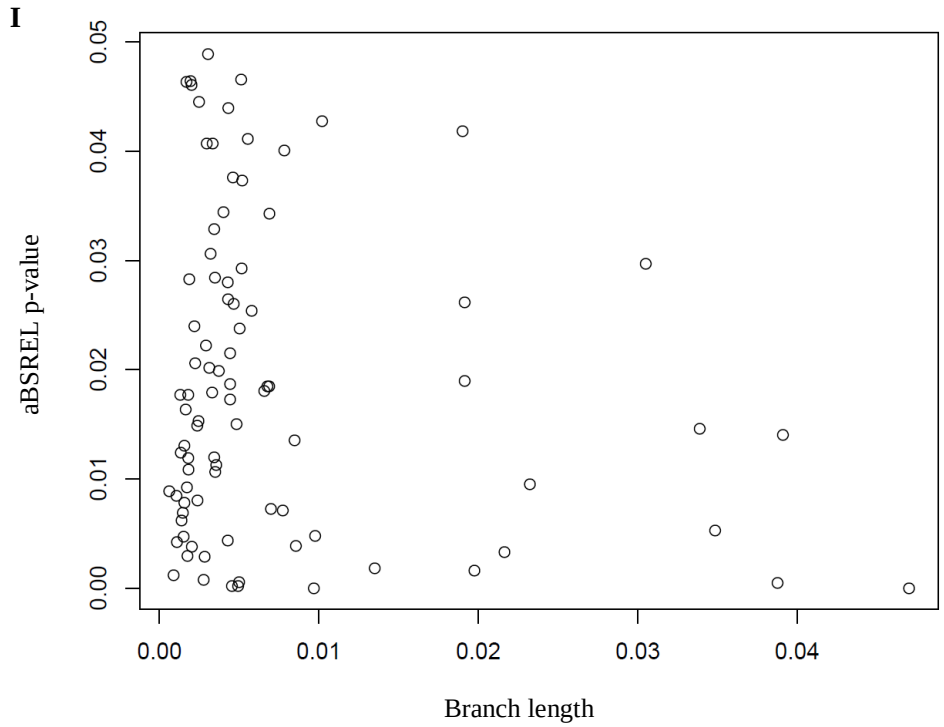
Genes	Study	Finding	Functional Category
per1b	Moreno, J., 2018	Signatures of positive selection	circadian rhythm
per2	Moreno, J., 2018	Signatures of positive selection	circadian rhythm
per3	Moreno, J., 2018	Signatures of positive selection	circadian rhythm
clocka	Moreno, J., 2018	Signatures of positive selection	circadian rhythm
clockb	Moreno, J., 2018	Signatures of positive selection	circadian rhythm
bmal2	Moreno, J., 2018	Signatures of positive selection	circadian rhythm
timeless	Moreno, J., 2018	Signatures of positive selection	cell stress response, circadian rhythm

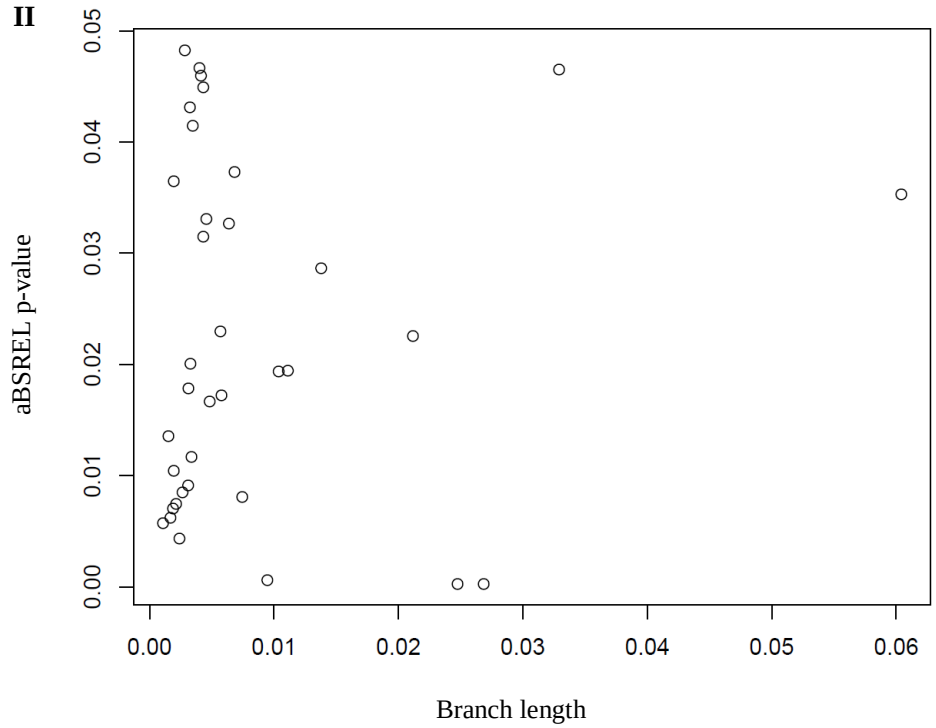
Supplementary Table 3.1 – Frequency and proportion of the 9 less frequent gene tree clusters (GTC) of Dataset D.

GTC	Frequency	Proportion
D7	481	0.06
D8	401	0.05
D9	337	0.04
D10	313	0.04
D11	283	0.04
D12	233	0.03
D13	221	0.03
D14	210	0.03
D15	140	0.02



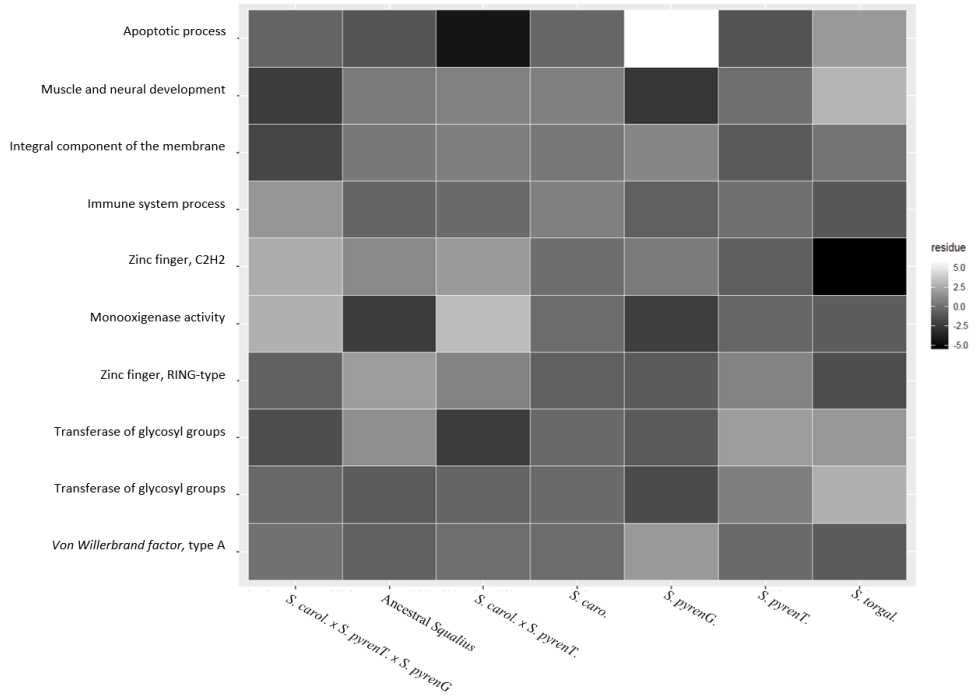
Supplementary Figure 3.1 - Median topology of the gene tree clusters (GTC) for the eight less frequent GTC's on Dataset D. S. pyrenaicus (G) refers to the Guadiana population and S. pyrenaicus (T) refers to the Tagus population of S. pyrenaicus.



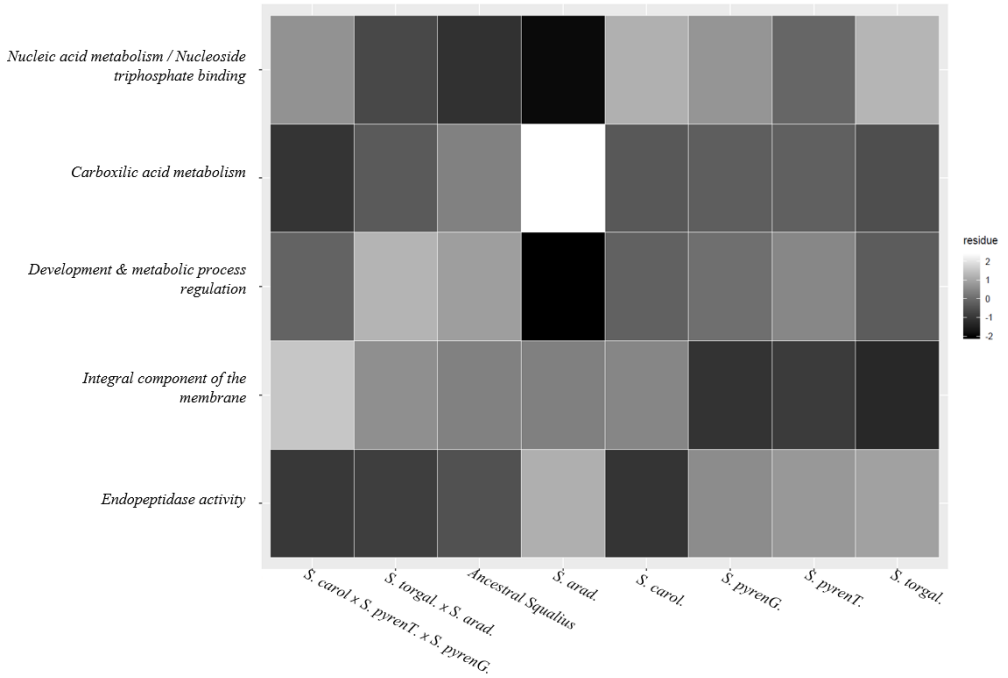


Supplementary Figure 3.2 - Plot of the branch lengths versus the respective aBSREL p-values of all the tip branches (extant species) with signatures of selection on I) Dataset C and II) Dataset D.

I

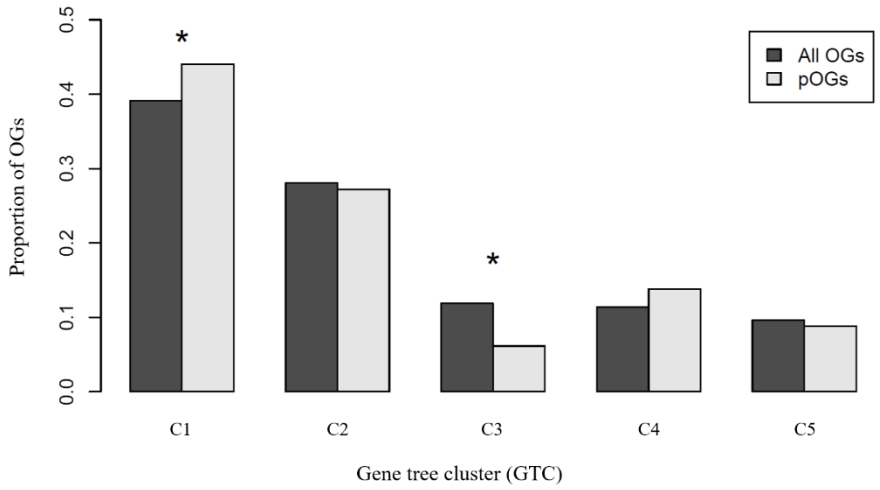


II

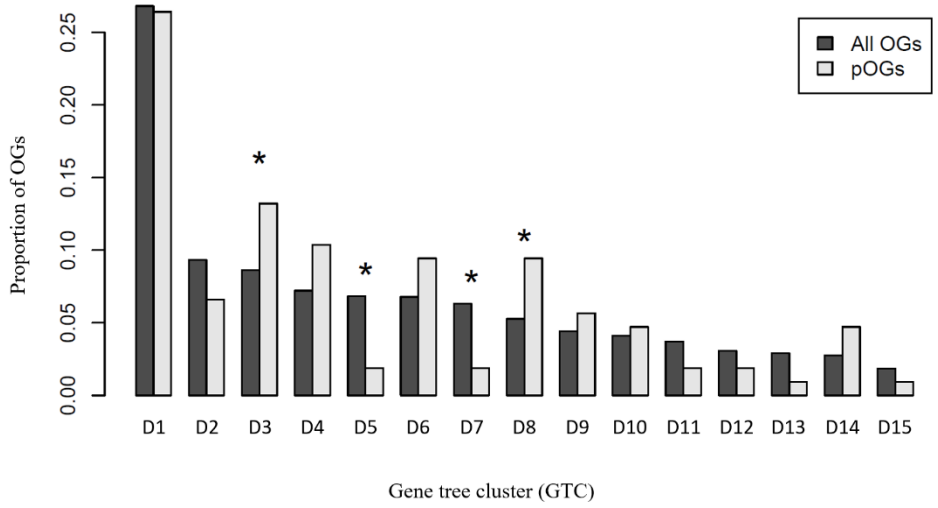


Supplementary Figure 3.3 – Heatmap of the residuals between the expected number of pOGs from each branch of the species tree in to each functional cluster and the observed number of pOGs from each branch on each functional cluster in I) Dataset C and II) Dataset D. The expected values were computed assuming that the two factors (functional cluster and branch of the gene tree) were independent. Branches with parentheses and several species names inside indicate the ancestral of those species. *S. carol.* – *S. carolitertii*, *S. pyrenT* – *S. pyrenaicus (Tagus)*, *S. pyrenG* – *S. pyrenaicus (Guadiana)*, *S. torgal* – *S. torgalensis*, *S. arad.* – *S. aradensis*, Ancestral *Squalius* – Ancestral branch of all *Squalius* species on that dataset, Others – Ancestral of all the *Squalius* and one of the outgroups.

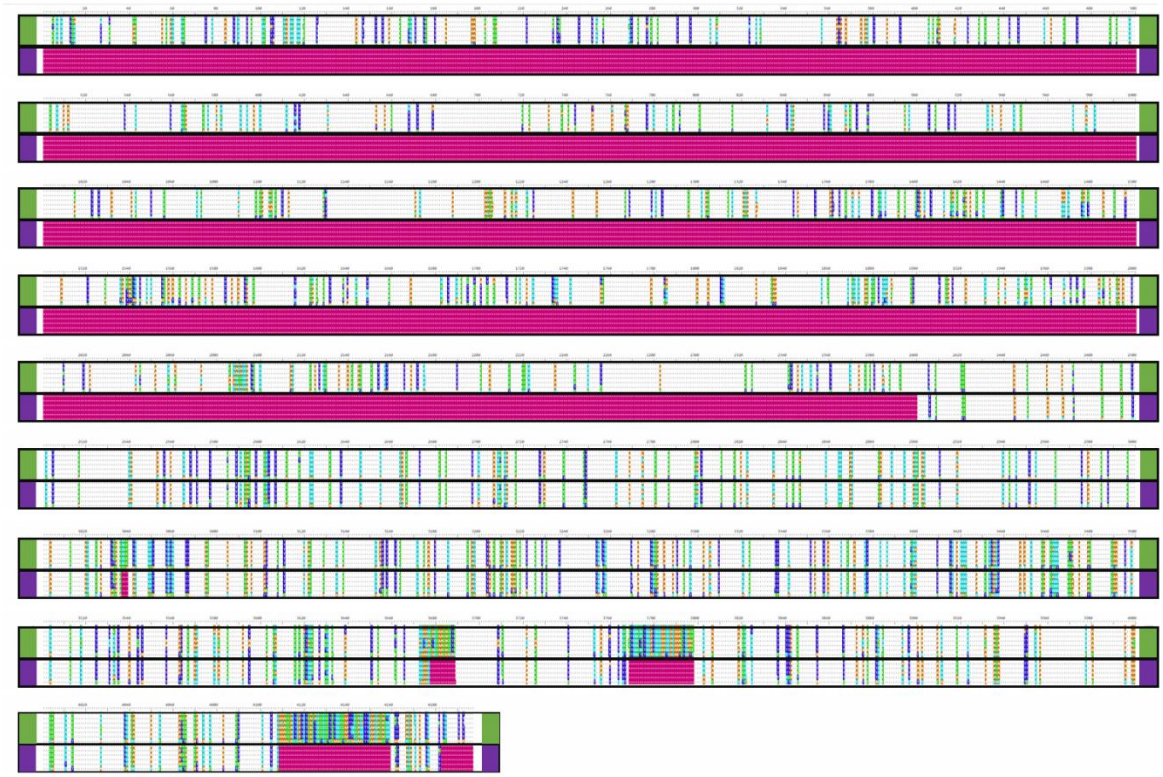
I



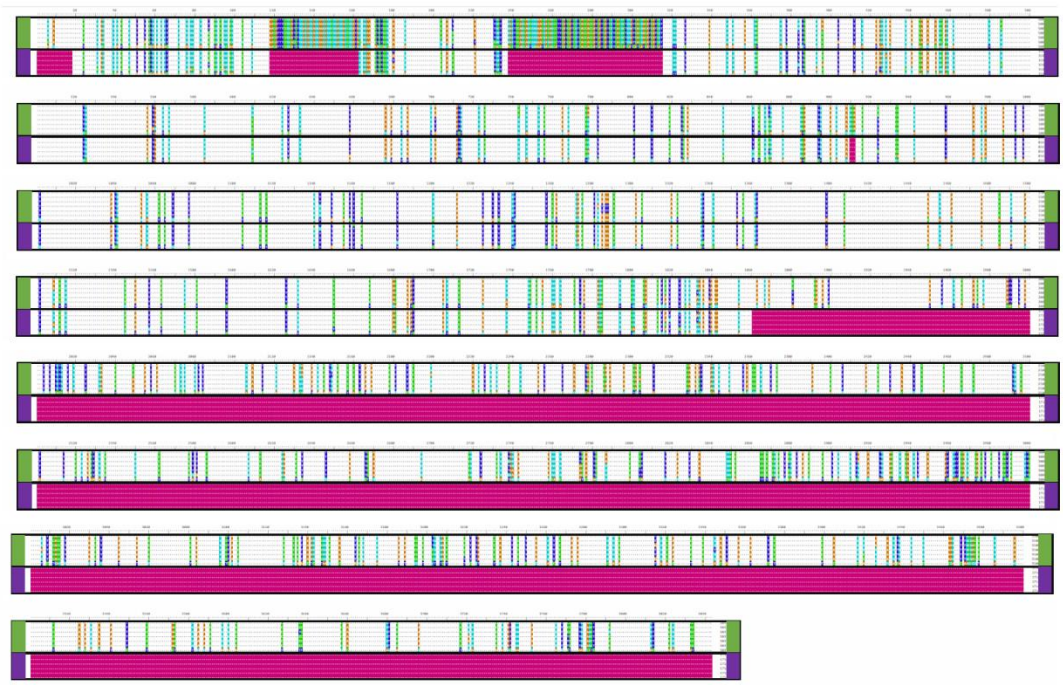
II



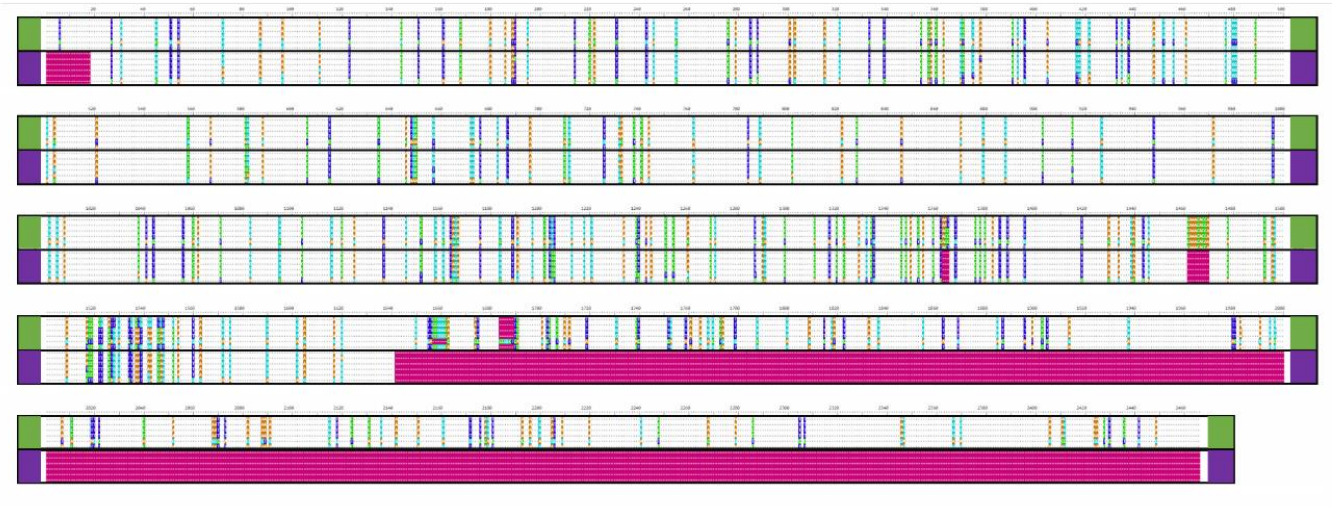
Supplementary Figure 3.4 – Comparison for I) Dataset C and II) Dataset D of the proportion of orthologous groups (All OGs) and proportion of orthologous groups under positive selection (pOGs) supporting each gene tree cluster (GTC). Asterisk indicate GTC's where the proportion of pOGs is significantly higher or lower than the proportion of OGs.



Supplementary Figure 3.5 – Comparison between Sanger (Moreno, J. 2018) and transcriptome (this study) sequences of the circadian rhythm gene Per2. White loci on the alignment represent conserved regions, coloured loci represent either polymorphisms or gaps on that position. Pink means that sequences has no information on that loci – either because the sequence is incomplete or because it is a gap. The first lines (delimited by green bars) correspond to the Sanger sequences, and the last lines (delimited by purple bars) correspond to the transcriptome sequences after our Pipeline 3. Results show that our alignments are missing a considerable fraction of the complete gene sequence.



Supplementary Figure 3.6 – Comparison between Sanger (Moreno, J. 2018) and transcriptome (this study) sequences of the circadian rhythm gene Per3. White loci on the alignment represent conserved regions, coloured loci represent either polymorphisms or gaps on that position. Pink means that sequences has no information on that loci – either because the sequence is incomplete or because it is a gap. The first lines (delimited by green bars) correspond to the sanger sequences, and the last lines (delimited by purple bars) correspond to the transcripome sequences after our Pipeline 3. Results show that our alignments are missing a considerable fraction of the complete gene sequence.



Supplementary Figure 3.7 – Comparison between Sanger (Moreno, J. 2018) and transcriptome (this study) sequences of the circadian rhythm gene Clockb. White loci on the alignment represent conserved regions, coloured loci represent either polymorphisms or gaps on that position. Pink means that sequences has no information on that loci – either because the sequence is incomplete or because it is a gap. The first lines (delimited by green bars) correspond to the sanger sequences, and the last lines (delimited by purple bars) correspond to the transcripome sequences after our Pipeline 3. Results show that our alignments are missing a considerable fraction of the complete gene sequence.

Supplementary Files

Additional supplementary material that could not be added to this document. It can be found on the following Dropbox link:

https://www.dropbox.com/sh/1v9s83bmqjmy4x1/AAD7dFA44e7nkttNiqOFHwiua?dl=0

Supplementary File 3.1 – Information on all the ortholog groups identified in Dataset C.

Supplementary File 3.2 – Information on all the ortholog groups identified in Dataset D.

Supplementary File 3.3 – Report of our misalignments trimming script for all the alignments of Dataset C.

Supplementary File 3.4 Report of our misalignments trimming script for all the alignments of Dataset D.

Supplementary File 3.5 – Summary of the aBSREL results for all the ortholog groups of Dataset C.

Supplementary File 3.6 – Summary of the aBSREL results for all the ortholog groups of Dataset D.

Supplementary File 3.7 – Information on all the ortholog groups with signatures of positive selection (pOG) on Dataset C.

Supplementary File 3.8 – Information on all the ortholog groups with signatures of positive selection (pOG) on Dataset D.

Supplementary File 3.9 – Results from DAVID's Functional Clustering Analysis for Dataset C.

Supplementary File 3.10 – Results from DAVID's Functional Clustering Analysis for Dataset C.

Supplementary File 3.11 – Temperature response target genes present on the list of ortholog groups and on the list of ortholog groups with signatures of selection on Dataset C and D.

Supplementary File 3. 12 – Circadian rhythm target genes present on the list of ortholog groups and on the list of ortholog groups with signatures of selection on Dataset C and D.

Supplementary Folder 2.1 – Custom scripts created for this work.