On the genetic basis of phenotypic diversity in teleost fish

Doctoral thesis for obtaining the
Academic degree
Doctor of Natural Sciences
(Dr. rer. nat.)

submitted by
Jan Gerwin

at the

Universität Konstantz

Faculty of Sciences
Department of Biology

Konstanz,
2022

Date of oral examination: 13.02.2023
First referee: Prof. Dr. Axel Meyer
Second referee: Prof. Dr. Catherine Peichel
Third referee: Prof. Dr. Christian Voolstra
Für Sophie.
Table of contents

Table of contents ............................................................................................... i

List of Figures ........................................................................................................ iv
  Main figures ........................................................................................................ iv
  Supplementary figures ....................................................................................... iv

List of Tables .......................................................................................................... vi
  Main tables .......................................................................................................... vi
  Supplementary tables ......................................................................................... vi

Summary ................................................................................................................. vii

Zusammenfassung ................................................................................................. x

General introduction ......................................................................................... 1
  Using next generation sequencing to investigate the genetic basis of phenotypic differentiation . 2
  Teleosts as model organisms to investigate the genetic basis of diversification .......... 3
  Chapter summary .............................................................................................. 5

Chapter 1  Of bars and stripes: a Malawi cichlid hybrid cross provides insights into
genetic modularity and evolution of modifier loci underlying color pattern diversification . 7

Abstract .............................................................................................................. 7

Introduction ........................................................................................................ 9

Materials and Methods .................................................................................... 12
  Experimental hybrid cross ............................................................................. 12
  Photographs and visual phenotyping .......................................................... 12
  Phenotyping using patternize ...................................................................... 13
  DNA isolation and RAD sequencing ............................................................. 14
  Sequence processing, marker selection and linkage map construction ......... 15
  QTL mapping .................................................................................................. 16

Results ................................................................................................................. 17
  Independent inheritance of horizontal stripes and vertical bars ..................... 17
  QTL analysis of stripe coverage and bar number suggests multigenic basis of horizontal stripes and vertical bars .......... 19
  Distinct QTL control anterior-posterior variation of horizontal stripes .......... 23
  Spatial analysis of vertical bars reveals additional major effect QTL .......... 25

Discussion .......................................................................................................... 27
  Modularity and independent evolution of stripe and bar patterns .................. 27
Chapter 2  

*Brain transcriptomics provide insights into short- and long-term regulation of aggressive behavior in the Siamese fighting fish, Betta splendens*  

Abstract ..................................................................................................................................... 46  
Introduction .............................................................................................................................. 48  
Materials and methods............................................................................................................................................. 52  
Animal experiments and dissection ..................................................................................................................... 52  
Behavioral scoring ........................................................................................................................................ 52  
RNA extraction ............................................................................................................................................. 53  
Library preparation, sequencing and read processing ....................................................................................... 53  
Differential gene expression analysis ............................................................................................................... 54  
Weighted correlation network analysis (WGCNA) .......................................................................................... 54  
Results ...................................................................................................................................... 56  
Behavioral differences between study species ............................................................................................... 56  
Transcriptomes of brain regions ....................................................................................................................... 56  
Transcriptional response to aggressive interaction .......................................................................................... 58  
Candidate genes and pathways ....................................................................................................................... 60  
Co-expression gene networks ......................................................................................................................... 68  
Discussion ...................................................................................................................................... 71  
Rapid transcriptional response in the brain of *B. splendens* .............................................................................. 71  
Short-term regulation of social behavior through changes in metabolism and endocrine system .......... 72  
Neuronal plasticity and long-term regulation of social behavior .................................................................. 74  
Conclusion............................................................................................................................................... 75  
Acknowledgements ..................................................................................................................................... 75  
Supplementary materials ................................................................................................................................. 76  

Chapter 3  

*Seeing red: evolution of cone opsin genes in the Siamese fighting fish, Betta splendens*  

Abstract ............................................................................................................................................... 83  
Introduction ............................................................................................................................................... 84
List of Figures

Main figures

**Figure 1.1** Diversity of melanic patterns in Lake Malawi haplochromine cichlid fishes. .............................................. 10

**Figure 1.2** Horizontal stripes in F\textsubscript{2} generation. ........................................................................................................ 18

**Figure 1.3** Vertical bars in F\textsubscript{2} generation. .................................................................................................................. 19

**Figure 1.4** QTL mapping of visually phenotyped stripe and bar patterns in Lake Malawi cichlids. .................. 21

**Figure 1.5** QTL mapping of stripe and bar patterns phenotyped using patternize. ................................................. 24

**Figure 2.1** Behavioral differences between the study species *B. simplex* and *B. splendens*. .................. 57

**Figure 2.2** Multidimensional scaling of transcriptome samples. ............................................................................. 59

**Figure 2.3** Up- and down-regulation of genes in different brain regions. ............................................................. 60

**Figure 2.4** Functional enrichment in differentially expressed genes of the telencephalon. .......................... 61

**Figure 2.5** Functional enrichment in differentially expressed genes of the diencephalon. .................................... 62

**Figure 2.6** Functional enrichment in genes of the mirror-specific modules of the diencephalon. ................. 67

**Figure 2.7** Functional enrichment in genes of the mirror-specific modules of the telencephalon. ................. 69

**Figure 3.1** Evolutionary history of opsin genes in the Anabantiformes. ................................................................. 93

**Figure 3.2** Synteny dot plot of genomic region harbouring *lws1a* and *lws1b*. ................................................................. 95

**Figure 3.3** Amino acid diversity of LWS between Betta and other teleosts and within Betta. ................. 97

**Figure 3.4** Ontogenetic change of relative expression of different opsin paralogs. ........................................ 100

Supplementary figures

**Figure S 1.1**: Stripe measurements and landmarks. ........................................................................................................ 39

**Figure S 1.2** Patternize PCA for parental species, DLS and MLS. .................................................................................. 40

**Figure S 1.3** Patternize PCA for parental species, vertical bars. .............................................................................. 41

**Figure S 1.4** Horizontal stripes and vertical bars in the F\textsubscript{1} generation. ............................................................. 42

**Figure S 1.5** Effect plots of QTL for dorsolateral stripe. .............................................................................................. 43

**Figure S 1.6** Effect plots of QTL for midlateral stripe. ................................................................................................. 44

**Figure S 1.7** Effect plots of QTL for full bars. ................................................................................................................. 45

**Figure S 2.1** Experimental setup. ................................................................................................................................. 76
Figure S 2.2 Venn diagrams showing overlap of up and downregulated genes in B. simplex. ....................... 77
Figure S 2.3 Venn diagrams showing overlap of up and downregulated genes in B. splendens. ................. 78
Figure S 2.4 Heat map comparing consensus and mirror modules in the telencephalon of B. simplex. ............... 79
Figure S 2.5 Heat map comparing consensus and mirror modules in the diencephalon of B. simplex. ............ 80
Figure S 2.6 Heat map comparing consensus and mirror modules in the telencephalon of B. splendens. ......... 81
Figure S 2.7 Heat map comparing consensus and mirror modules in the diencephalon of B. splendens. ........... 82
Figure S 3.1 Phylogenetic reconstruction and recombination break points of lws. ........................................ 108
Figure S 3.2 Phylogenetic reconstruction of rh2 and sws. ................................................................................. 109
Figure S 3.3 Synteny dot plot of lws locus. ........................................................................................................ 110
Figure S 3.4 Phylogenetic reconstruction and recombination break points of rh2 ............................................. 111
Figure S 3.5 Amino acid diversity of RH2 between Betta and other teleosts and within Betta. .................. 112
Figure S 3.6 Amino acid diversity of SWS2 between Betta and other teleosts and within Betta. .............. 113
List of Tables

Main tables

Table 2.1 Top six differentially expressed genes in B. simplex. ................................................................. 64
Table 2.2 Top six differentially expressed genes the diencephalon of B. splendens. ........................................ 65
Table 2.3 Top six differentially expressed genes the telencephalon of B. splendens. ........................................ 66
Table 3.1 Estimated maximal absorption of lws genes in B. splendens. .......................................................... 96

Supplementary tables

Table S 1.1: Summary of multiple QTL models of visual measurements ......................................................... 37
Table S 1.2: Summary of multiple QTL models of patternize analysis .......................................................... 38
Table S 3.1 Estimated maximal absorption of lws genes in B. splendens......................................................... 114
Summary

The origin of phenotypic diversity in closely related species remains one of the great mysteries of biology. Exploring the genetic mechanisms underlying this phenotypic diversity is one of the main endeavors of evolutionary biology. The recent advances in sequencing technologies allow the investigation of the genetic basis of phenotypic diversity in a wide range of organisms. The infraclass of teleosts is a lineage that gave rise to a multitude of adaptive radiations, of which many represent great model systems to study the genetic basis of morphology and behavior. In this thesis, I combined a number of different sequencing approaches to investigate the genetic basis of different morphological features and behaviors using two different teleost species radiations: the haplochromine cichlids of East African Great Malawi and the South-East Asian fighting fishes.

In chapter 1, we investigated the genetic basis of the two most common melanic patterns found in African cichlids: horizontal stripes and vertical bars. Using a hybrid cross between the striped species *Pseudotropheus cyaneorhabdos* and the barred species *Chindongo demasoni*, we conducted quantitative trait locus (QTL) mapping, to identify genomic locations underlying the formation of the two color patterns. The distribution of phenotypes in our hybrid cross showed, that bars and stripes are inherited as independent modules. The results of the QLT analysis revealed, that in Lake Malawi cichlids one large effect QTL encompassing the gene *agrp2* affects the formation of horizontal stripes next to a number of smaller effect QTL. While most parts of the horizontal stripe pattern share a common genetic basis, a few of the identified QTL are only responsible for the formation of specific parts of the stripe pattern. The analysis of vertical bars revealed a high number of small effect QTL underlying the formation of the color pattern. We discuss the role of hybridization in the diversification of melanic color patterns in Lake Malawi cichlids regarding the modularity of stripes and bars and the independent inheritance of some parts of the horizontal stripe patterns.

Our work in chapter 2 focuses on the genetic basis of aggressive behavior in two fighting fish species of the genus *Betta*. Using brain transcriptomes of the highly aggressive species *B. splendens*
and the less aggressive species *B. simplex*, we investigated the species-specific transcriptional responses to an aggressive interaction, to detect expressional differences that might underlie the behavioral differences of the two studied species. Differential gene expression analysis revealed a much stronger transcriptional response in the aggressive species *B. splendens* when compared to *B. simplex*. The transcriptional response was strongest in the diencephalon and telencephalon of *B. splendens*. A closer look at the functions of the genes showed, that *B. splendens* showed differential expression of genes involved in hormonal regulation and regulation of the carbohydrate metabolism, indicating changes in the short-term regulation of aggressive behavior in this species. Both species showed differential expression of genes that may have functions in the long-term regulation of aggressive behavior by the formation of new neurons and neuronal connections. Weighted correlation network analysis revealed that genes involved in long-term regulation of social behavior were integrated into new co-expression networks in the telencephalon of both species. Our results indicate a more fine-tuned short-term regulation of aggressive behavior in the aggressive species *B. splendens*, highlighting the importance of regulation of behavior in an aggressive environment.

In Chapter 3, we investigated the visual system of the Siamese fighting fish *B. splendens* with a focus on the cone opsin gene *lws* of which five copies were found in the genome of *B. splendens*. Using whole genome sequences of nine different anabantoid species we explored the evolutionary history of cone opsins in the Anabantiformes. Our analysis shows, that only species of the genus *Betta* have five copies of *lws* and that at least two duplication events in the *Betta* lineage lead to the expansion of this cone opsin class. The analysis of the amino acid sequences of the different *lws* paralogs revealed functional differentiation between the different paralogs within *Betta*. Using eye transcriptomes of *B. splendens*, we demonstrated expression shifts between paralogs of all cone opsin classes throughout ontogeny. Expression of *lws* showed patterns of temporal collinearity, which might have facilitated the expansion and functional diversification of *lws* genes in *Betta*.

The three chapters emphasize the importance of using a variation of approaches to investigate the
genetic basis of different phenotypes in different teleost model systems in order to get a comprehensive understanding of the fast evolution of phenotypic diversity.
Zusammenfassung


In Kapitel 1 sind wir den genetischen Grundlagen von horizontalen Streifen und vertikalen Balken, welche die beiden häufigsten, melanistischen Farbmuster afrikanischer Buntbarsche sind, auf den Grund gegangen. Mit Hilfe eines Hybridstammes der horizontal gemusterten Art *Pseudotropheus cyaneorhabdos* und der vertikal gemusterten Art *Chindongo demasoni* haben wir eine Quantitative Trait Locus (QTL) Analyse durchgeführt, um genomische Regionen zu identifizieren, die die Bildung der beiden Farbmuster beeinflussen. Die Verteilung der Phänotypen in den Hybriden zeigt, dass Balken und Streifen als unabhängige Module vererbt werden. Die Ergebnisse der QTL Analyse zeigen, dass ein genomischer Locus, der das Gen *agrp2* beherbergt, großen Einfluss auf die Bildung horizontaler Streifen nimmt. Außerdem sind mehrere QTL mit kleinerem Effekt an der Bildung beteiligt. Während die meisten QTL die Bildung des gesamten Farbmusters beeinflussen, sind ein paar QTL nur an der Bildung bestimmter Teile des horizontalen Streifenmusters beteiligt. Die QTL Analyse des vertikalen Farbmusters zeigt, dass viele QTL mit kleinem Effekt die Bildung von vertikalen Balken beeinflussen. Wir diskutieren die mögliche Rolle von Hybridisierung bei der Diversifizierung melanischer Farbmuster in Buntbarschen des Malawisees und beziehen uns dabei auf die von uns aufgedeckte Modularität von...


Die drei Kapitel dieser Dissertation heben hervor, wie wichtig es ist, verschiedene Methoden und verschiedene Modelorganismen zu kombinieren, um die genetische Grundlage verschiedener Phänotypen zu erforschen und letztendlich ein umfassendes Verständnis von der schnellen Evolution phänotypischer Diversität zu erhalten.
General introduction

Explaining the rise of phenotypic diversity in closely related species constitutes one of the central problems in evolutionary biology, as it contradicts the strict dogma of slow and gradual change during evolutionary processes (Gould 1980). While it was possible to investigate the ultimate causes (evolutionary history and adaptive value) of phenotypic diversity for decades, only the relatively recent rise of new molecular technologies allowed for an extensive investigation of the proximate causes (cellular, molecular and genetic basis) of the evolution of phenotypes. Some of the most famous textbook examples of evolutionary biology nicely illustrate how the advance in technology facilitated the exploration of proximate causes of fast phenotypic changes. For instance, after the first discoveries of the melanic form of the peppered moth (*Biston betularia*) in its native range in the UK in the mid-19th century and the first proposal of the adaptive advantage in the polluted forests of the industrialized UK (Tutt 1896), Kettlewell (1955) was the first to empirically demonstrate the adaptive value of the dark color morph. The proximate cause, the genetic variant underlying the dark color morph, was only discovered over 50 years later through means of whole genome sequencing and association mapping (van’t Hof *et al.* 2011; van’t Hof *et al.* 2016). Similarly, the adaptive value of the beak shapes of Darwin’s finches of the Galapagos islands in exploiting different food sources has been described decades ago (Grant *et al.* 1976), but the underlying molecular mechanisms leading to differentiation of beak in the different species were only discovered much later through the use of different technologies like in-situ hybridization (Abzhanov *et al.* 2004), RNA-sequencing (Abzhanov *et al.* 2006) and whole genome sequencing (Lamichhaney *et al.* 2015).
Using next generation sequencing to investigate the genetic basis of phenotypic differentiation

As these two examples demonstrate, the investigation of the genetic bases of phenotypes was immensely facilitated through the emergence of next generation sequencing technologies allowing for high-throughput sequencing of both DNA and RNA. The combination of different sequencing techniques and different analytical methods allows for a multitude of approaches to investigate the genetics of phenotypes using nucleotide sequences. While some of these approaches make use of whole genome sequences, others only rely on sequencing of parts of the organisms DNA or tissue specific RNA.

One of the most popular methods to find genomic regions responsible for phenotypic variation is quantitative trait locus (QTL) mapping. This approach, making use of one of the fundamental phenomena in genetics, recombination, allows the efficient determination of genomic regions causing phenotypical differences between closely related species or different populations of the same species (Tanksley & Nelson 1995). Early QTL studies used relatively few genetic markers to identify large genomic regions encompassing the genetic variants of interest (Paterson et al. 1988; Paterson et al. 1989). With the decreasing cost for sequencing, it became possible to increase the number of individuals and markers, allowing for the investigation of genetically complex, morphological (Henning et al. 2017; Gerwin et al. 2021; Navon et al. 2021) and behavioral phenotypes (Greenwood et al. 2016; Bendesky et al. 2017; Feller et al. 2020), even identifying loci that only have small effects on the formation of the phenotype. The main draw backs of QTL mapping are that only closely related species can be investigated, as hybridization into the second offspring generation must be possible (Tanksley 1993), and the relatively low resolution of the results. In most cases, QTL mapping indicates the presence of the causative genetic polymorphism in a large genomic region, harboring multiple genes (van’t Hof et al. 2011; Henning et al. 2014). In these cases, it can be helpful to complement the genetic mapping approach with methods involving the determination of expression levels of different genes via transcriptomics (Fruciano et al. 2016).
Transcriptomic analysis can be used to measure gene expression differences between the different phenotypes of interest. In contrast to the genetic mapping approach, transcriptomes reveal actual biological differences between phenotypes that go beyond the mere DNA sequence difference. While expression differences found through RNA sequencing might reveal actual biological differences, the underlying genetic difference (e.g. sequence differences in regulatory sequence or coding sequence of a transcription factor) might stay undetected. Furthermore, expression differences correlating with phenotypic differences do not necessarily cause the observed phenotypic difference. Combining both genetic tools will lead to a powerful tool of finding the underlying genetic variation causing phenotypic variation. In the presented thesis, I will use both techniques to study the genomic underpinnings of coloration, aggressive behavior and the visual system in various teleost fish species demonstrating the adaptability and strength these methods have in modern evolutionary biology.

Teleosts as model organisms to investigate the genetic basis of diversification

With more than 26,000 described species the infraclass of teleost fishes makes up about half of all described vertebrate species. Further, teleosts harbor countless radiations of closely related species showing phenotypic variation. Some of these radiations constitute well-suited model systems for the investigation of rapid evolution of striking phenotypic differentiation among closely related species. Exploring the genetics of phenotypic differentiation in radiations at different stages of divergence can be used to answer a multitude of different questions.

In this thesis, I made use of two different teleost radiations, addressing questions revolving around phenotypic diversification within these radiations. The first radiation, the haplochromine cichlids of the African Great Lakes Victoria and Malawi, represent one of the most popular model system used in evolutionary biology (Chapter 1). These lakes were colonized independently followed by the evolution of adaptive radiations (Salzburger et al. 2005), leading to both immense phenotypic diversification within the lakes (Danley & Kocher 2001) and convergent evolution between the radiations (Meyer 1993). Cichlids of Lake Malawi diversified along many different phenotypic axis including trophic morphology (Albertson et al. 2003; Hulsey et al. 2019), body shape (Husemann et al.
2014, 2017), breeding behavior (York et al. 2015; York et al. 2018) and coloration (Allender et al. 2003; Kocher 2004; Gerwin et al. 2021) allowing to investigate the evolutionary mechanisms underlying a plethora of different phenotypes. In addition, the different colonization histories of Lake Malawi (radiation started about 800,000 years ago (Brawand et al. 2014)) and Lake Victoria (radiation started 10,000 – 100,000 years ago (Verheyen et al. 2003; Elmer et al. 2009)) give us the opportunity to look at different stages of adaptive radiation and the evolution of phenotypic variation. This setup makes it possible to explore the molecular mechanisms of rapid phenotypic changes, but also compare this to how evolution affects the genetic basis of traits over longer periods of time (Kratochwil et al. 2018; Gerwin et al. 2021).

The second radiation used for the studies in this thesis are species of the genus Betta (Chapters 2 and 3). The more than 70 described species (Linke 2014) of the genus occur in different fresh water habitats of South-East Asia. In comparison to African cichlids, the genus Betta is relatively old and diverged around 20 million years ago (Rüber et al. 2006). Still, hybridization within some of the different breeding groups of the genus is possible (Kwon et al. 2022). Comparable to East African cichlids, species of the genus Betta diverged along different phenotypical axis including body size (Linke 2014), coloration (Linke 2014; Kwon et al. 2022), breeding behavior (Rüber et al. 2004) and aggressive behavior (Verbeek et al. 2008) making them suitable to investigate the genetic basis of these phenotypes. Despite their rich phenotypic diversity, most studies working on this genus only included the most famous species of the genus, the Siamese fighting fish, Betta splendens, and almost exclusively explored the aggressive behavior of this species. B. splendens has been a popular subject in ethology for decades (Baenninger 1966; Simpson 1968), as it reliably displays aggressive behaviors and is easy to keep and breed in the lab. While most studies on B. splendens investigated some aspect of their social behavior (Evans 1985; Bronstein 1994; McGregor et al. 2001; Dziewczynski et al. 2005; Dziewczynski et al. 2009), with the recent publication of high quality reference genomes (Fan et al. 2018; Kwon et al. 2022) exploring the genetics of behavior (Vu et al. 2020; Vu et al. 2022) and morphology (Wang et al. 2021; Kwon et al. 2022) of this species became more popular. The
combination of a well-established model organism for the investigation of aggressive behavior and the phylogenetic background with many closely related species, that show lower levels of aggression, make the genus *Betta* a perfect model system to study the genetic underpinnings of aggressive behavior.

**Chapter summary**

This thesis consists of three chapters investigating the genetic basis of phenotypic diversity in closely related teleost species. Working on species from two different teleost radiations, the haplochromine cichlids of the East-African Great Lake Malawi and the South-East Asian fighting fishes, I investigated the underlying evolutionary and genetic mechanisms influencing morphological and behavioral differentiation using a combination of several modern sequencing techniques.

In Chapter 1, we investigated the genetic basis of the two most common melanic color patterns found in African cichlids: horizontal stripes and vertical bars. Using a hybrid cross between the striped species *Pseudotropheus cyaneorhabdos* and the barred species *Chindongo demasoni*, we conducted a quantitative trait locus (QTL) analysis to identify genomic regions underlying the formation of either trait. The analysis of the genetic basis of horizontal stripes revealed that horizontal stripes are an oligogenic trait affected by one large effect QTL harboring the gene *agrp2* that has previously been linked to the formation of stripe patterns in African cichlids (Kratochwil *et al.* 2018). While the two parts of the stripe pattern mostly share a genetic basis, the formation of some portions of the color pattern is partially affected by different genomic loci, in contrast to previous results from Lake Victoria cichlids, where horizontal stripes are a monogenic trait. The analysis of the genetic basis of vertical bars revealed that bars are a polygenic trait affected by many genomic loci of small effect size, already indicated by the fact that none of the 230 individuals of the F2 generation used in this study lacked the bar pattern. We discuss the potential role of hybridization on the diversification of color patterns in Lake Malawi cichlids taking into account the multigenic basis of both color patterns.

Chapter 2 concentrates on the genetic basis of aggressive behavior in the fighting fishes of the genus *Betta*. Using a behavioral assay, individuals of the highly aggressive species *B. splendens* and the
less aggressive species *B. simplex* were exposed to their mirror image for either 10 or 30 minutes. After exposure, the brains were removed and transcriptomes of four different brain regions (telencephalon, diencephalon, optic lobes, cerebellum) were analyzed. Using the transcriptomes, we could compare gene expression differences between treatment and control individuals to investigate the transcriptional response to an aggressive interaction in both species. Our analysis of differential expression revealed a concerted transcriptional reaction in *B. splendens*, while *B. simplex* showed little transcriptional changes. Functional enrichment analysis revealed, that differentially expressed genes in *B. splendens* may influence the short-term regulation of social behavior in an aggressive interaction. Both species also showed enrichment of genes involved in the long-term regulation of aggressive behavior. Weighted correlation network analysis revealed that genes involved in neuronal plasticity and hence the long-term regulation of social behavior were introduced into new co-expression networks. Our results underline the important role of precise regulation of costly aggressive behavior and demonstrate the complexity of the evolution of social behavior.

In Chapter 3, we characterized the visual system of *B. splendens* and its close relatives with a focus on the red-sensitive cone opsin *lws*, of which five paralogs can be found in the genome of *B. splendens*. Using whole genome sequences of nine different anabantoid species, we analyzed the evolutionary history of cone opsin genes in this order. We could demonstrate that five copies of *lws* can only be found in the genomes of *Betta* and that at least two tandem duplication events in the lineage leading to *Betta* lead to the expansion of *lws* copy number. An analysis of the amino acid sequences of the *lws* paralogs in *Betta* revealed diversification of the sequence at four of the seven transmembrane domains of *lws*, indicating functional differences between the paralogs. Finally, transcriptomic analysis of eye tissue revealed expression switches between paralogs for all cone opsin classes throughout ontogeny. We highlight that the temporal collinearity of *lws* expression might have facilitated the expansion and diversification of *lws* in *Betta*. Our work emphasizes the important role of gene duplication in the evolution of phenotypic diversity.
Chapter 1
Of bars and stripes: a Malawi cichlid hybrid cross provides insights into genetic modularity and evolution of modifier loci underlying color pattern diversification

Jan Gerwin, Sabine Urban, Axel Meyer and Claudius F. Kratochwil
Published in Molecular Ecology

Abstract

Understanding the origins of phenotypic diversity among closely related species remains an important largely unsolved question in evolutionary biology. With over 800 species, Lake Malawi haplochromine cichlid fishes are a prominent example of extremely fast evolution of diversity including variation in coloration. Previously, a single major effect gene, agrp2 (asip2b), has been linked to evolutionary losses and gains of horizontal stripe patterns in cichlids, but it remains unknown what causes more fine-scale variation in the number and continuity of the stripes. Also, the genetic basis of the most common color pattern in African cichlids, vertical bars, and potential interactions between the two color patterns remain unknown. Based on a hybrid cross of the horizontally striped Lake Malawi cichlid Pseudotropheus cyaneorhabdos and the vertically barred species Chindongo demasoni we investigated the genetic basis of both color patterns. The distribution of phenotypes in the F2 generation of the cross indicates that horizontal stripes and vertical bars are independently inherited patterns that are caused by two sets of genetic modules. While horizontal stripes are largely controlled by few major effect loci, vertical bars are a highly polygenic trait. Horizontal stripes show substantial variation in the F2 generation that, interestingly, resemble naturally occurring phenotypes found in other Lake Malawi cichlid species. Quantitative trait loci (QTL) mapping of this cross reveals known (agrp2) and unknown
loci underlying horizontal stripe patterns. These findings provide novel insights into the incremental fine-tuning of an adaptive trait that diversified through the evolution of additional modifier loci.
Chapter 1

Introduction

Understanding the origin of phenotypic diversity during the formation of adaptive radiations remains a question that is only now being addressed with genomic and forward genetic approaches. Adaptive radiations are often characterized by shared genetic variation and ongoing hybridization that can both facilitate (Meier et al. 2017; Svardal et al. 2020) but in some cases also counteract morphological diversification (i.e. if species collapse; Mallet 2007). The cichlid fishes inhabiting the East African Great Lake Malawi offer an example of an extremely young adaptive radiation (2–4 million years) with >800 genetically very similar (Meyer et al. 1990; Verheyen et al. 2003; Malinsky; et al. 2018), but phenotypically highly diverse species. Axes of diversification include among others trophic morphology, body shape and coloration. Beyond diverse nuptial colorations, Lake Malawi cichlids greatly vary in melanic patterns (i.e., based on melanophores, melanin-bearing pigment cells (Fig. 1.1) (Hendrick et al. 2019; Liang et al. 2020; Liang; et al. 2020)). The two most common color patterns in African cichlids are vertical bars and horizontal stripes. Both patterns show meristic variation (i.e., number of stripes/bars) and differ in width, continuity, orientation and contrast. Horizontal stripe patterns consist of one or two distinct stripes, the dorsolateral stripe (DLS) and midlateral stripe (MLS) (Fig. 1.1 A) (Seehausen et al. 1999). Vertical bar patterns usually consist of six or more bands covering the distance between the dorsal fin and the ventral part of the body (Fig. 1.1 C, F). Although vertical bars and horizontal stripes were likely already present at the origin of the radiation (Barlow 2000), we find a variety of combinations and modifications of these patterns. Therefore, one of the questions we want to address is how the diversity of melanic patterns in Lake Malawi cichlids (Fig. 1.1) could have evolved within such a short evolutionary time.

Previous work suggests that the presence (or absence) of horizontal stripe patterns is controlled by regulatory alleles at a single major effect locus that affect agouti related peptide 2 (agrp2) expression and thereby stripe absence (high agrp2 expression) or presence (low agrp2 expression) (Kratochwil et al. 2018). In Lake Victoria cichlids, the youngest major species flock that is less than 0.1
million years old (Meyer et al. 1990; Verheyen et al. 2003) stripes follow a Mendelian inheritance pattern. In contrast, a hybrid cross between a striped and a non-striped Lake Malawi cichlid species shows that in Lake Malawi only around 50 percent of the variance is explained by this locus suggesting a more complex genetic basis (Kratochwil et al. 2018). Hence, it is possible that additional modifier loci were already present prior to or evolved within the older Lake Malawi radiation. These additional loci might then also explain the greater variation of stripe patterns found in Lake Malawi compared to the much younger Lake Victoria haplochromine cichlid radiation (Fig. 1.1). Alternatively, diversity in color patterns could be a result of reshuffling of ancestral standing genetic variation that entered Lake Malawi with the colonizing individuals and/or subsequent hybridization (Brawand et al. 2014; Svardal et al. 2020). The genetic basis for vertical bar patterns in cichlids is still unknown, although we gained some initial insights into their development and the genes that might play a role in their development (Hendrick et al. 2019; Liang et al. 2020; Liang; et al. 2020).

Figure 1.1 Diversity of melanic patterns in Lake Malawi haplochromine cichlid fishes. Each panel shows a photograph of a cichlid species from Lake Malawi and a schematic representation of its horizontal stripes and vertical bars. Stripes vary in several aspects including continuity (e.g., B and E), number (e.g., A and D), orientation (e.g., B and D), and contrast (e.g., B and E), bars mainly in number (e.g., C and F) and contrast (e.g., B and C). Abbreviations: DLS, dorsolateral stripe. MLS, midlateral stripe. A striking observation is the seemingly
independent evolution of horizontal stripe and vertical bar patterns. The majority of cichlid fish species of the East African radiations (probably over 800 species in the three adaptive radiations of Lake Victoria, Malawi and Tanganyika and the riverine, non-endemic cichlids of East Africa) only displays either horizontal stripes (about 1/3 of species) or vertical bars (most of the remaining 2/3 of species) (Urban et al. submitted). In the few species that show both color patterns one of them is usually fainter than the other one (e.g., Fig. 1.1 B). This might be ultimately driven by a functional mismatch of the traits (as previously described for other traits, e.g., Arnegard et al. (2014)), as the two traits might constitute alternative and mutually exclusive strategies to hide from predators or prey (Seehausen 1999) or mutually exclusive developmental trajectories. To better understand the evolution of these two traits in the haplochromine cichlid radiations of East Africa, here with a focus on Lake Malawi cichlids, it is important to examine how independent the two traits are in terms of their genetic basis and how this might have facilitated or constrained color pattern divergence.

In an effort to gain novel insights into the genetic basis and evolution of stripe and bar patterns in cichlid fishes we performed a hybrid cross between the striped species *Pseudotropheus cyaneorhabdos* and the barred species *Chindongo demasoni* (previously: *Pseudotropheus demasoni*) from Lake Malawi (the same cross as used for single-locus association mapping in Kratochwil et al. (2018)). One notable feature of cichlid fishes is that many can be bred easily in the laboratory, and that fertile crosses are possible among many species, in particular among the haplochromine cichlids of East Africa that make up the adaptive radiations of Lakes Malawi and Victoria. This permits breeding experiments such as the one we performed for this study. This cross allowed us to (i) identify the genetic architecture of both traits, (ii) to test if vertical bars and horizontal stripes are independently inherited traits and (iii) to identify quantitative trait loci (QTL) that underlie the variation of stripe patterns found in Lake Malawi cichlid species.
Materials and Methods

Experimental hybrid cross
Six *C. demasoni* females and one *P. cyaneorhabdosa* male were set up in a 200 L aquarium until one of the females carried eggs in her mouth. The female was isolated in a separate tank until the offspring was released. The F₁ offspring was split into two identical 100 L tanks and raised until sexual maturity (between 6 and 9 months). The F₁ individuals were allowed to mate randomly to obtain individuals of the F₂ generation. Eggs were removed from the F₁ females after 6 to 10 days after fertilization and were reared in egg tumblers (Zissaqua Zet-65). F₂ individuals were raised for 12 months in 100 L aquaria in groups of ca. 20 individuals. Offspring of 33 broods was used in the final QTL analysis. We did not control for maternity nor paternity as the two pigmentation patterns of the parental species do not differ between the sexes. In this study we used individuals of the same hybrid cross as in Kratochwil *et al.* (2018) but at a later life stage that permitted a more refined phenotyping (12 months of age in this study compared to 3 months in the previous study).

Photographs and visual phenotyping
F₂ individuals were photographed from left and right side at the age of 12 months in a 25 cm x 10 cm x 5 cm aquarium that was placed in a white photo shooting tent using a digital camera (Canon EOS 7D, 100 mm lens). The dorsolateral (DLS) and midlateral stripes (MLS) (Fig. 1.1 A) were analyzed separately. Phenotyping of stripes was performed in three different ways. First, stripes were measured as a binary trait, where a completely or partially present stripe was scored as 1 and a completely missing stripe was scored as 0 (only used for Chi-square test). Second, stripes were measured as an ordinal phenotype. Here, stripes covering the whole anterior-posterior range of the potentially striped region were scored as 2, stripes that only covered parts of that region were scored as 1 and completely missing stripes were scored as 0 (only used to give an overview of the phenotypic distribution). Third, the proportion of the anterior-posterior range of the potentially striped region (the region that is
striped in the striped parental species *P. cyaneorhabdos* covered by melanic pigmentation was measured (used for QTL and Figure 1.2). Here, the stripes could potentially be scored in a range between 0% and 100%. To exclude the effect of vertical bars for the continuous stripe measurement, we measured both the potentially and actually striped regions between the vertical bars only (for a more detailed description see the supplementary text and Fig. S1.1). Vertical bars were counted and classified into two categories: full bars (bars covering the whole dorso-ventral range as seen in the parental species *C. demasoni*) and partial bars (bars not covering the whole range). Three different phenotype measures where obtained: number of full bars, number of partial bars and total bar number. All phenotypic measurements described above were scored on the left and right side. The average of both sides was used as the phenotype for QTL mapping.

Phenotyping using *patternize*

Even though the measurement of stripe coverage takes into account the continuous distribution of stripe phenotypes in the F2 generation of our cross it is still limited regarding one important factor: it does not account for location-specific losses/gains of the stripe patterns. The R package *patternize* (Van Belleghem *et al.* 2018) enables extraction of color patterns observed in photographs of different individuals of the F2 generation and the parental species and to compare these color patterns in a principal component analysis (PCA). This analysis can be conducted for any given part of the color pattern separately. Using the function *patLanRGB* the photographs were aligned according to fixed landmarks (Fig. S1.1) previously determined on each photograph. Next, color information was extracted from each photograph in windows of a predefined size (each of these windows represents a square of pixels). The color information was translated into a binary matrix (1 = color present, 0 = color absent), that was then used for PC analysis. The color information of all photographs used in our analysis was transformed to grey values (black and white photographs) using Adobe Photoshop CS6. Furthermore, the brightness levels of the photographs were adjusted using the levels tool in Photoshop. To correct for the different intensity of the color pattern in different individuals, the
darkest part of the color patterns was set to be the darkest part of the photograph and the lightest part of the fish to be the lightest part of the photograph. For our analysis, we created color value matrices for each photograph of the left side of the F2 individuals and for eight individuals of each of the parental species using the `patLanRGB` function (with parameters: RGB = c(0,0,0), colOffset = 0.35, res = 200). To create a morphospace of extreme values we first analyzed the matrices of the parental individuals in a PCA (Fig. S1.2, Fig. S1.3). Using the R function `predict` we then projected the color value matrices of the F2 individuals into the parental morphospace. The resulting PC values were then used for QTL mapping.

For our analysis, we did not analyze the entire color pattern, but limited the PC analysis to the coloration in our regions of interest. The `maskOutline` function allows the user to limit the analysis to a specific region. We wanted to specifically analyze the regions covered by the mid- and dorsolateral stripes and vertical bars. Furthermore, we wanted to analyze different parts of the stripes separately.

A major obstacle in the analysis was that horizontal stripes and vertical bars overlap in F2 individuals that show both color patterns. As the color of stripes and bars is the same, it is not possible to reliably distinguish between black coloration contributed by a bar or by a stripe in the overlapping regions. For this reason, we limited the analysis with `patternize` to those regions where color patterns in the parental species do not overlap. Consequently, the analysis for the presence and absence of bars was limited to the region between the mid- and dorsolateral stripe and the region ventrally of the midlateral stripe. The analysis of the mid- and dorsolateral stripe was only conducted in the regions between the vertical bars.

DNA isolation and RAD sequencing

All F2 individuals were fin clipped directly after taking the photographs for phenotyping. Fin clips were stored in 100% ethanol at -20°C until DNA isolation. About 0.25 cm² of fin tissue were used for DNA isolation using the Tissue DNA Purification Mini Spin Column Kit (Genaxxon Bioscience) following the manufacturer’s instructions. Before the extraction, the tissue was incubated in TE buffer (500mM Tris,
20mM EDTA, 10mM NaCl, pH 9.0) for one hour at room temperature.

The isolated DNA was used to prepare double digest restriction site associated DNA (ddRAD) sequencing libraries (Peterson et al. 2012; Franchini et al. 2017). Briefly, double restriction enzyme digestion was performed on 1µg of sample DNA using the enzymes PstI and MspI (New England Biolabs). Then P1 and P2 adapters were ligated to the digested DNA using T4 ligase (New England Biolabs). The Pippin Prep (Sage Science) was used to select DNA fragments of ca 450-550 bp and DNA libraries of 279 F2 individuals were pooled in five sequencing pools. Each library pool contained one library of each of the parental individuals. Paired end sequencing (150 bp) was performed using the Illumina HiSeq X Ten platform (one library per lane) at the Beijing Genomics Institute, Hong Kong.

Sequence processing, marker selection and linkage map construction
The sequences were processed using the STACKS pipeline (version 2.41) (Catchen et al. 2013). Briefly, PCR duplicates were removed using the clone_filter program. Using the inner and outer barcodes, the sequences were assigned to the corresponding individuals. Non-paired reads were excluded from further analysis. Next, the reads were aligned to the *Maylandia zebra* genome (Conte & Kocher 2015) using the Burrows-Wheeler Aligner (Li & Durbin 2009) and sorted using samtools (Li et al. 2009). The program ref_map.pl (part of STACKS) was used for SNP calling and the populations program was run to generate population-level statistics.

Next, we used the program qtools.py (Kautt et al. 2020) for marker selection. Uninformative markers, markers missing in the parental samples and markers that are heterozygous in the parental samples were removed. The remaining 6322 loci were further filtered and all loci that were present in less than 95% of the samples were removed.

For linkage map construction, the resulting file was imported into the program JoinMap 4 (Build: 4may07.4oct06), where identical loci and loci that significantly deviated from Hardy-Weinberg-Equilibrium were excluded from further analysis using the respective functions. The linkage map was constructed using the Haldane’s mapping function under default settings.
QTL mapping

QTL mapping was performed using the R package qtl (Broman et al. 2003). After loading the phenotype-genotype data frame using the function read.cross, the function jittermap was used to adjust the position of overlapping markers. Next, the functions calc.genoprob and sim.geno were used to calculate genotype probabilities (step = 0.1, error.prob = 0.00001) and to simulate genotypes between the observed markers (step = 0.1, n.draws = 64). QTL mapping was performed using the Haley-Knott method (Haley & Knott 1992) in the function scanone. To identify significant QTL the genome wide significance threshold was calculated using the scanone function (n.perm = 1000, 95th percentile LOD as threshold).

The function scanone works under the assumption of a single QTL and ignores the possibility of multiple QTL and possible epistatic effects between the loci. To find all loci associated with the phenotypes of interest and to investigate their potential interactions, we searched for further suggestive QTL by looking for those QTL only exceeding the chromosome wide threshold. Here, the significance threshold was calculated for each chromosome using the scanone function, but this time the desired chromosome was specified (n.perm = 1000, chr = x, 90th percentile LOD as threshold to be sure to find all significant QTL). All identified QTL (exceeding genome and chromosome wide thresholds) were combined using the makeqtl function. The created QTL object was then analyzed using the fitqtl function that calculates the percentage of variation explained (PVE) by the whole model and by each QTL. Non-significant QTL were dropped one by one and the analysis was repeated until all added QTL showed a significant effect in the multiple-QTL model.

QTL intervals were determined using the function bayesint of the R package qtl. The markers closest to the limits of the QTL intervals were used to identify a corresponding genomic region in the reference genome of the closely related species Maylandia zebra (Conte & Kocher 2015). Names of known genes in the genomic regions were extracted using the biomart package in R. We identified genes that might be involved in coloration by filtering out those genes that had the terms “pigment” and “melan” as a part of their Gene Ontology (GO) description.
Results

Independent inheritance of horizontal stripes and vertical bars

The first part of the study addresses the question if the two different color patterns are inherited independently by conducting a hybrid cross. We analyzed the presence of vertical bars and horizontal stripes in 230 adult individuals of the second offspring (F$_2$) generation of our hybrid cross between the striped species *Pseudotropheus cyaneorhabdos* and the barred species *Chindongo demasoni*. Both vertical bars and horizontal stripes can be found in individuals of the F$_1$ (Fig. S1.4) and the F$_2$ generation (Fig. 1.2). The average number of bars in the F$_2$ generation is ~8.5 bars (average between both sides of the fish) with a maximum of 10.5 bars and a minimum of one bar (Fig. 1.3). Individuals of the barred parental species *C. demasoni* usually show either six or seven vertical bars. Not all bars in individuals of the F$_2$ generation fully cover the flank from dorsal to ventral as they do in *C. demasoni* (Fig. 1.1 F). Interestingly the inheritance of vertical bars significantly deviates from a 3:1 Mendelian ratio (Chi-square test, $\chi^2$=76.6, df=1, P<0.001) with all individuals showing some kind of bar pattern, suggesting a highly polygenic basis of the trait.

The distribution of horizontal stripes shows variation between the dorsolateral stripe (DLS, Fig. 1.1 D) and the midlateral stripe (MLS). The analysis of the ordinal measurements shows that 16.25% of the F$_2$ individuals lack the DLS completely, while only 1.25% of them do not show any parts of the MLS. Also, here both traits, DLS (Chi-square test, $\chi^2$=87.7, df=1, P<0.001) and MLS (Chi-square test, $\chi^2$=560.7, df=1, P<0.001) deviate from a 3:1 ratio that would have been expected if horizontal stripes had a simple Mendelian basis as shown before for the extremely young radiation of haplochromine cichlids of Lake Victoria (Henning *et al.* 2014). As stripes vary substantially in their continuity, we additionally scored the stripe phenotype as a continuous trait by calculating the proportion of the anterior-posterior range of the potentially striped region that is covered by melanic pigmentation (stripe coverage). To account for confounding effects of the vertical bars that cross the stripe region, we only measured the striped regions between the vertical bars. For the DLS, F$_2$ individuals showed a stripe coverage between 22 and 100%, for the MLS between 50 and 100% (Fig. 1.2 G–I).
Figure 1.2 Horizontal stripes in F₂ generation. Individuals of the F₂ generation show a range of expression of the horizontal stripe pattern: (A) complete pattern, (B, E) parts of the dorso- and midlateral stripe, (C) parts of the midlateral stripe, (D) midlateral stripe only, (F) no horizontal stripes. G shows the correlation of the continuous measurement of the DLS and the MLS. Panels H and I show the correlation between the number of full bars and the continuous measurements of the DLS and the MLS, respectively. The blue lines in panels G, H and I show linear regressions between the traits for visualization of the direction of the correlation. The results of Spearman’s Rank Correlation test are shown in panels G, H and I. Abbreviations: DLS, dorsolateral stripe. MLS, midlateral stripe.

To test for an association between the presence and absence of horizontal stripes and vertical bars we used the Fisher’s exact test. We found no association between the presence of the MLS and bars (Fisher’s exact test, $P = 1$) and between the presence of the DLS and bars (Fisher’s exact test, $P = 1$). For a more detailed analyses we calculated Spearman’s rank correlation coefficients to test for the correlation between the number of full bars and stripe coverage of the DLS and the MLS. We did not detect a significant correlation between the number of full bars and DLS ($\rho = -0.12, P = 0.065$) (Fig. 1.2 H). In contrast, we found a significant negative correlation between the number of full bars and
MLS (\(\rho = -0.24, P = 0.0002\)) (Fig. 1.2 I). The weak correlation between the MLS and the number of full bars generally indicates an independent genetic basis of both phenotypes, but hints at the possible sharing of small effect loci. However, continuous measurements of DLS and MLS show a strong positive correlation (\(\rho = 0.72, P < 0.001\)) (Fig. 1.2 G), possibly suggesting a shared genetic basis of the two horizontal stripes.

Figure 1.3 Vertical bars in F\(_2\) generation. Individuals of the F\(_2\) generation show different numbers of vertical bars. A shows the distribution of the number of full bars in the F\(_2\) generation. The grey area in A shows the phenotypic range of the barred parental species \(C.\) demasoni. Panels B – F show individuals of the F\(_2\) generation with different bar numbers and schematic illustrations of their bar patterns.

QTL analysis of stripe coverage and bar number suggests multigenic basis of horizontal stripes and vertical bars

To further investigate the more complex genetic basis of stripe and bar patterns in Lake Malawi cichlids as well as to identify the loci associated with the traits, we performed QTL mapping. As we observed
substantial variation in the phenotypes of both stripes we used measured stripe coverage as a more objective quantification instead of binary or ordinal measurements. To do so, we constructed a linkage map using quadRAD markers from 230 F$_2$ individuals of the hybrid cross. The final linkage map contains 2770 markers arranged in 22 linkage groups corresponding to the 22 chromosomes of the high-quality reference genome of the closely related species *Maylandia zebra* (Conte & Kocher 2015). The total length of the linkage map is 1297.8 cM with a mean marker distance of 0.47 cM. To identify all loci associated with vertical bar and horizontal stripe formation we tested for QTL exceeding the genome wide significance threshold (large effect QTL) as well as the chromosome-wide significance threshold (which we will refer to as suggestive QTL).

For the DLS, we found three QTL with log of the odds ratio (LOD) scores exceeding the genome wide LOD threshold on chromosomes 2 (LOD 5.6), 5 (LOD 4.6) and 18 (LOD 20.5) (Fig. 1.4 A). Additionally, we could identify four suggestive QTL on chromosomes 7, 17, 22 and 23. A multiple QTL model including all seven identified QTL explains 56.8% of the variation of the phenotype. The QTL on chromosome 18 that harbors the previously described major effect locus *agrp2* (Kratochwil et al. 2018) has the largest effect (24.4 percent variance explained (PVE)), followed by the QTL on chromosomes 2, 5 and 17 (~4 PVE). The remaining three QTL on chromosomes 7, 22 and 23 explain very little of the observed variation (<3 PVE) (Table S1.1, Fig. S1.5).

For the MLS we found three significant QTL on chromosomes 2 (LOD 4.8), 10 (LOD 4) and 18 (LOD 12.2) (Fig. 1.4 B). Four suggestive QTL were found on chromosomes 3, 5, 15 and 17. A multiple QTL model including all identified QTL explains 42.2% of the observed variation of the MLS. Again, the QTL on chromosome 18 has the largest effect (14.1 PVE), while the QTL on chromosomes 2 and 10 show moderate effects (~3 PVE). The remaining four QTL explain only a small fraction of the variation (<3 PVE) (Table S1.1, Fig. S1.6). The PVE of the QTL found on chromosome 18 is lower than previously reported with a single-locus model and using ordinal scoring of the MLS (52.7 PVE; Kratochwil et al. 2018). This difference is however mainly due to the phenotyping, as ordinal scoring of the MLS gives more comparable PVE for the QTL on chromosome 18 (31.8 PVE).
Figure 1.4 QTL mapping of visually phenotyped stripe and bar patterns in Lake Malawi cichlids. LOD scores across all 22 chromosomes for (A) the dorsolateral stripe, (B) the midlateral stripe and (C) complete vertical bars. D shows LOD scores across the chromosomes with QTL exceeding the genome wide significance threshold for all three phenotypes. The grey area in D indicates the position of the agrp2 locus in the linkage map. Generally, grey dashed lines indicate genome wide significance thresholds and black arrows indicate QTL exceeding genome wide significance thresholds. Green arrows indicate suggestive QTL. The effect plots show the relationship between the genotype at the QTL on chromosome 18 and (E) the dorsolateral stripe, (F) the midlateral stripe and
(G) the number of full bars. Points show the mean of the phenotype, error bars indicate the standard error. Abbreviations: LOD, logarithm of the odds. BB, homozygous for allele of the barred parental species. SB, heterozygous. SS, homozygous for allele of the striped parental species. DLS, dorsolateral stripe. MLS, midlateral stripe.

Next, we screened for QTL associated with the vertical bar patterns, a trait that had not been genetically mapped before. A limitation caused by the distribution of bar phenotypes in the F2 mapping panel is, that we could not map the presence/absence of the trait, as all F2 individuals had at least one bar (Fig. 1.3). Also, due to meristic variation, a similar measurement of coverage as for the stripe phenotype is not possible. We hence focused on the number of bars as focal trait (Fig. 1.3 A). When analyzing the total number of bars (including complete bars and bars that do span the entire sides of the fish from dorsal to ventral) we could not identify any significant QTL. A more specific analysis of the number of complete bars revealed a single QTL with a LOD score of 5.3 (Fig. 1.4 C). We found five suggestive QTL on chromosomes 5, 7, 13, 14 and 18. A multiple QTL model including all six QTL explains 32% of the variation in the phenotype. In this case, the QTL located on chromosome 2 shows the largest effect on the phenotype (6.8 PVE). All other QTL explain 5% or less of the variation (Table S1.1, Fig. S1.7).

In summary, both of the horizontal stripes share the QTL with the largest effects on chromosomes 2 and 18 (the latter harboring the previously associated agrp2 locus) and two additional loci with smaller effects on chromosomes 5 and 17. The overlap of the QTL therefore likely contributes to the strong correlation of the DLS and MLS (Fig. 1.2 G). Moreover, each stripe shows individual and non-overlapping QTL with smaller effect sizes (chromosomes 7, 22 and 23 for the DLS; chromosomes 3, 10 and 15 for the MLS) that might account for differences in the patterns. The analysis of the number of complete bars revealed one large effect QTL on chromosome 2 in a different location than the QTL associated with horizontal stripes suggesting that different genetic modules control the two melanin patterns. Three suggestive loci exclusive for vertical bars can be found on chromosomes 5, 13 and 14 and QTL that are shared between vertical bars and horizontal stripes could be found on chromosomes 7 and 18 (Fig. 1.4 C). The multiple QTL models explain a substantially higher percentage of variation.
for the two horizontal stripes (56.8 and 42.2 PVE for DLS and MLS, respectively) when compared to vertical bars (32.9 PVE).

Distinct QTL control anterior-posterior variation of horizontal stripes

A striking observation from the phenotypic distribution of the $F_2$ individuals is the occurrence of intermediate phenotypes (Fig. 1.2 B-E) including more spotted or discontinuous stripe patterns that resemble naturally occurring phenotypes of other species (Fig. 1.1 C, D, E). This suggests that certain loci might regulate specific parts of the stripes and act on top of the previously described role of agrp2 (Chr. 18) that merely acts as a more general inhibitor and on-off-switch of the whole pattern.

To test this hypothesis, we used a more fine-grained analysis of the spatial pattern using the R-package patternize (Van Belleghem et al. 2018). The software allows automatic alignment of images of the $F_2$ and parental individuals based on manually set morphometric landmarks followed by automatized extraction of color information. Through principal component analysis of the extracted color information we generated PC values for each $F_2$ individual that could then be used for QTL mapping. This approach allowed us to analyze specific homologous regions of the respective color patterns.

In a first step we compared this approach to the manual measurement (Fig. 1.4). The results were similar and revealed two QTL with LODs exceeding the genome wide threshold on chromosomes 2 (LOD 7.2) and 18 (LOD 4.2) for the dorsolateral stripe. Contrary to the analysis of the visual phenotyping, the QTL on chromosome 2 shows the largest effect. Two suggestive QTL were found on chromosomes 22 and 23 (Fig. 1.5 A). A multiple QTL model including all four identified QTL explains 27.1% of the observed variation. The QTL on chromosome 2 has the largest effect on the phenotype (11 PVE), followed by the QTL on chromosome 18 (4.3 PVE). Each of the remaining four QTL explains less than 4% of the variation (Table S1.2). The peaks on Chr. 2, 18, 22 and 23 are shared between the manual and patternize analysis.
The analysis of the midlateral stripe revealed a single significant QTL on chromosome 18 (LOD 6.2). Additionally, we found four suggestive QTL on chromosomes 1, 7, 8 and 17 (Fig. 1.5 B). The multiple QTL model explains 28% of the observed variation. Again, the QTL on chromosome 18 shows the largest effect (10 PVE), while each of the remaining QTL explains less than 4% of the variation (Table S1.2). The peaks on Chr. 17 and 18 are shared between the manual and patternize analysis (Fig. 1.4).

Figure 1.5 QTL mapping of stripe and bar patterns phenotyped using patternize. LOD scores across all 22 chromosomes for (A) the DLS, (B) the MLS, (C) the anterior DLS, (D) the anterior MLS, (E) the posterior DLS, (F) the posterior MLS, (G) the dorsal part of vertical bars, (H) the ventral part of the vertical bars. Generally, grey dashed lines indicate genome wide significance thresholds. Black arrows indicate QTL exceeding genome wide significance thresholds. Green arrows indicate QTL exceeding chromosome wide significance thresholds.
To investigate the genetic basis of variation in the horizontal stripe pattern, we split each stripe into an anterior and a posterior part and performed separate QTL mapping for each of these parts. The detailed analysis of the DLS revealed that a large effect QTL on chromosome 2 is responsible for most of the variation observed in the anterior part (11.5 PVE) (Fig. 1.5 C), while the large effect QTL on chromosome 18 is responsible for the variation observed in the posterior part of the DLS (7.3 PVE) (Fig. 1.5 E). Similarly, the spatial analysis of the MLS showed that the largest effect QTL on chromosome 18 is responsible for 12% of the variation of the posterior part of the stripe, too (Fig. 1.5 D). No large effect QTL were found for the anterior part of the MLS (Fig. 1.5 F). Hence, we suggest that specific loci might be responsible for driving variation in specific parts of the stripe pattern.

Spatial analysis of vertical bars reveals additional major effect QTL

Many species show variation in the intensity of melanic pigmentation along the dorso-ventral axis, which includes bar patterns that often fade more ventrally — a phenotype that is also apparent in a substantial number of F2 individuals (Fig. 1.2 B, E). Therefore, we investigated if there is a QTL associated with this dorso-ventral patterning of vertical bars using patternize (Fig. 1.5 G and H). To avoid interference with the horizontal stripes we limited the analysis to the dorsal (between the DLS and the MLS) and the ventral region (ventral of the MLS). The analysis of the dorsal region revealed a single significant QTL (LOD 4.3) on chromosome 17. Seven other suggestive QTL were found on chromosomes 2, 3, 5, 11, 14, 18 and 20 (Fig. 1.5 G). A multiple QTL model including all identified QTL explains 36.3% of the observed variation of the bar phenotype. The QTL on chromosome 17 shows the largest effect on bar presence (8.5 PVE), while the remaining QTL explain less than 4.2% of the variation. The analysis of the ventral region revealed one QTL with a LOD value exceeding the genome wide significance threshold on chromosome 7 (LOD 9.6). We found three suggestive QTL on chromosomes 17, 18 and 22 (Fig. 1.5 H). A multiple QTL model including all identified QTL explains 27.7% of the observed phenotypic variation. The QTL on chromosome 7 has the largest effect (14.9 PVE). Each of the remaining QTL explain less than 5% of the observed variation (Table S1.2). Therefore,
similar to the results for the stripe patterns the in-depth analysis of the bar patterns suggests a spatial
control of the dorso-ventral pattern.
Discussion

We investigated the genetic basis of the two most common color patterns in African cichlids, horizontal stripes and vertical bars, using a hybrid cross between the Lake Malawi species *P. cyaneorhabdos* and *C. demasoni*. We identified a multitude of genomic loci explaining the variation of the two color patterns that we can see between the two parental species. Our results provide insights into the modularity of color patterns as well as how more fine-scale variation of these patterns might be caused. Lastly, these results in combination with previous work gives, despite being speculative, insights into how the dynamics of color pattern evolution might be constrained and facilitated by their genetic architecture.

Modularity and independent evolution of stripe and bar patterns

Our first observation is that bar and stripe patterns are inherited independently. This is demonstrated by the independent assortment of the two traits in our hybrid cross. In case of a shared genetic basis of the two color patterns, the traits would show a strong negative correlation in the F₂ generation as different alleles at the shared color pattern loci would only allow the formation of either pattern. Furthermore, the QTL mapping does not show substantial overlap between the large effect QTL of bar and stripe patterns. These results suggest that the formation of vertical bars and horizontal stripes does not share the same genetic basis and that the traits thus constitute variational and developmental modules (sensu Wagner *et al.* (2007)), allowing independent evolution, loss or gain of one color pattern without affecting the other color pattern. Clearly, both patterns are shaped by variation in number and properties (e.g., pigment content) of the same cell types (melanophores in the dark regions, and xanthophores and iridophores in the regions in between). Yet the coordination (and the loci that underlie this coordination) of the pigment cells that are shaping bar and stripe patterns seem to be largely independent as the two phenotypes freely segregate in the F₂ generation providing no evidence for a shared genetic basis. This is further supported by previous results from a CRISPR-Cas9 knockout
of the major effect gene agrp2 that specifically affects stripe but not bar formation (Kratochwil et al. 2018). The fact that only rarely species display both stripes and bars could therefore be explained by a functional mismatch and not by any sort of developmental constraint or pleiotropic effect (as individuals with both patterns are very common in the F2 panel (Fig. 1.2)). A plausible explanation for the fact that both traits rarely occur in the same species (with exceptions, as for example multiple species in the genus Julidochromis, Neolamprologus buescheri, Haplochromis sauvagei, etc.) is that both traits constitute alternative but not compatible dazzle camouflage or communication strategies (the barring pattern becomes more pronounced in courting males in particular) that have adaptive optima in species with specific behavior (e.g., swimming speed), morphology (e.g., body shape) and environmental preferences (e.g., rocky, sandy, vegetated, open-water environments) (Seehausen et al. 1999). However, more robust evidence for the adaptive function of the two color patterns is still missing and needs further investigation. Due to the mostly uncoupled genetic basis of both traits, selection can effectively act on each trait separately. A similar example for such functional and genetic decoupling that greatly facilitated rates of diversification in cichlid fishes are pharyngeal and oral jaws (Hulsey et al. 2006; Parsons et al. 2012; Hulsey et al. 2016; Ahi et al. 2019b). We observed a weak negative correlation between the number of full bars and the MLS (Fig. 1.2 I), that is probably caused by the shared QTL on chromosome 18 (Figure 1.4). This QTL has the largest effect on the presence of both of the horizontal stripes (PVE 14 for the MLS), but it only has slight effects the presence of full bars (3.7 PVE), causing a weak correlation of the phenotypes but still granting a mostly independent inheritance of the traits.

The genetic basis of horizontal stripes: from major effect loci to incremental changes Previous results have demonstrated a simple Mendelian basis for stripe patterns in Lake Victoria cichlids that are caused by cis-regulatory mutations of the agrp2 gene and also identified the agrp2 locus on chromosome 18 — using a single locus test — as a major effect locus in Lake Malawi (Henning et al. 2014; Kratochwil et al. 2018). Here, using a genome-wide QTL search we confirm that agrp2 is
indeed the major effect locus underlying the presence of stripe patterns. On top of this, we have identified several loci that influence stripe patterns globally (Fig. 1.4) and spatially (Fig. 1.5). The fact that the identified QTL only explain around 50% of the observed variation (57% for the DLS, 42% for the MLS) indicates that there might be non-genetic factors that affect color pattern formation. Interestingly, by comparison with the Lake Victoria cross between the striped species *Haplochromis sauvagei* and the non-striped species *Pundamilia nyererei* (Henning et al. 2014), our Lake Malawi cross shows substantially more variation in the F_2_ generation. This includes variation in the continuity, but also in the number of stripes. And while there is a strong positive correlation between the presence and relative coverage of the DLS and the MLS (Fig. 1.2 G), several individuals (~22%) had only one of the stripes. Interestingly, it was always the DLS that was missing (no individual only lacked the MLS), suggesting that one or more underlying alleles are specifically affecting the DLS, while others affect both. Indeed, QTL analysis of the visually phenotyped DLS and the MLS revealed that two major effect loci (Chromosomes 2 and 18) were shared between the two stripes, while one major effect QTL was exclusively associated with each of the stripes (chromosome 5 for DLS, chromosome 10 for MLS) (Fig. 1.4).

Stripe patterns in the F_2_ panel did not only show meristic variation (i.e., in the number of stripes), but also variation in their continuity with some individuals having a more spotted or interrupted stripe and others only showing stripes in the anterior or posterior flank. Interestingly, many of these phenotypes (Fig. 1.2) were never seen in individuals of the parental species but had striking resemblance to stripe phenotypes of other cichlid species found in Lake Malawi (Fig. 1.2). For example, we can find phenotypes that are alike the stripe patterns of *Protomelas annectens* (DLS absent, MLS present; Fig. 1.1 D), *Cyrtocara moorii* (one anterior portion of the DLS, two posterior portions of the MLS; Fig. 1.1 E) or *Pseudotropheus crabro* (DLS absent, MLS present with small posterior gap; Fig. 1.1 C).

In the context of previous work these new results allow for interesting speculation on the evolutionary history of the loci underlying stripe formation in the haplochromine radiations of Lake
Malawi and Lake Victoria. Based on these results we propose two interpretations explaining the
differences of the genetic basis of gain and loss of horizontal stripes we found between the radiations
of Lakes Malawi and Victoria. The genetic basis of stripes in Lake Victoria cichlids is well explained by
regulatory evolution of two divergent haplotypes of a single intronic interval that evolved prior to the
radiation and seems to be responsible for most of the variation in stripe patterns across the ~500
species of this radiation (Kratochwil et al. 2018; Urban et al. 2020). The simple, monogenic basis of
horizontal stripes in Lake Victoria cichlids could be therefore the result of a bottleneck during the
colonization of Lake Victoria (i.e., other stripe-affecting alleles did not enter the Lake Victoria radiation
with the colonizers). The trait would have thereby evolved from being an oligogenic trait to a
monogenic trait. Another potential explanation is that the trait was initially a monogenic trait and that
this basis was maintained in Lake Victoria, but that additional stripe alleles only evolved in Lake Malawi.
This second hypothesis is supported by previous results suggesting that the stripe-affecting alleles from
Lake Victoria are older than the radiation itself and that stripe-associated alleles evolved de novo
within Lake Malawi (Urban et al. 2020). The evolution of more loci affecting the formation of horizontal
stripes in Lake Malawi might have been possible due to the older age of this radiation when compared
to the Lake Victoria radiation (10,000–100,000 years for Lake Victoria (Meyer et al. 1990; Seehausen
2002; Verheyen et al. 2003; Elmer et al. 2009; Wagner et al. 2013); 800,000 years for Lake Malawi
(Meyer et al. 1990; Brawand et al. 2014). In summary, stripes would have therefore continued to be a
monogenic trait in Lake Victoria, while in Lake Malawi new modifier loci have evolved leading to a
transition from a monogenic to an oligogenic trait. This hypothesis is supported by findings from
studies on the genetic basis of color patterns in the genus Danio. Here, it was shown that the color
pattern differences between a horizontally striped species (D. quagga) and a spotted species (D.
kyathit) has a complex genetic basis, even though transitions from the striped phenotype to the
spotted phenotype could be achieved by loss of function mutations in single genes in D. rerio
(McCluskey et al. 2021). This example illustrates that phenotypic differences between divergent
lineages are likely to be based on multiple genes even though the same phenotypic differences can
potentially be caused by a simple genetic change, supporting the accumulation of additional modifier loci during the course of evolution.

Alternatively to the evolution of novel stripe modifying alleles within Lake Malawi, it is also possible that the evolution of stripe patterns in Lake Malawi was solely driven by selection on pre-existing genetic variation (including variation that was introduced by introgression and/or hybridization in ancestral lineages) (Loh et al. 2013; Svardal et al. 2020). Most of the ancestral, riverine species including e.g., *Astatotilapia burtoni* and *Astatoreochromis alluaudi* have an interrupted stripe pattern that is much less pronounced than the stripes of most Malawi species and resembles some of the F$_2$ individuals (Fig. 1.2 B). Accordingly, the stripe phenotype diversity among cichlids in Lake Malawi might have been caused by incomplete lineage sorting of alleles (or even adaptive introgression) (Malinsky; et al. 2018; Svardal et al. 2020) influencing stripe pattern formation (even if the phenotypes were not present ancestrally because of low allele frequencies) and selection acting on the diverse phenotypes caused by these different allele combinations (including the strongly striped species like *P. cyaneorhabdos*). This would be comparable to e.g., the evolution of extreme body shapes from standing genetic variation via polygenic selection, as for example observed in cichlid crater lake radiations (Malinsky et al. 2015; Kautt et al. 2020). Furthermore, it would support the possible role of ancient variation in the formation of phenotypic diversity during rapid adaptation (Barrett & Schluter 2008; Brawand et al. 2014; Meier et al. 2017; Irisarri et al. 2018; Svardal et al. 2020). Both mechanisms (evolution of novel modifier alleles and evolution of novel allele combinations) or a combination of both might therefore have contributed to the diversity of stripe pattern phenotypes that makes up the radiation of cichlids in Lake Malawi.

The higher number of QTL in our Malawi cichlid hybrid cross therefore might suggest an explanation for the higher phenotypic diversity in stripe patterns among Lake Malawi cichlids, as compared to Lake Victoria cichlids, that include variation in the number, orientation, continuity and contrast of the stripes. Our cross provides experimental evidence that different combinations of alleles from only one striped and one non-striped species can indeed create a variety of naturally occurring
phenotypes different from those observed in the parents (Fig. 1.1). This supports the conclusion that novel combinations of alleles driven by hybridization and/or introgression could in fact have facilitated this diversification process in color patterns (as it did here in our hybridization experiment in the laboratory). It has been suggested before that hybridization played a key role in the evolution of the cichlid radiations of Lakes Victoria (Meier et al. 2017), Tanganyika (Salzburger et al. 2002; Koblmuller et al. 2007; Irisarri et al. 2018) and Malawi (Joyce et al. 2011; Genner & Turner 2012; Svardal et al. 2020) as it allows the reassembly of variants (Marques et al. 2019) that might ultimately shape the evolution of traits as we suggest here for color patterns.

Polygenic basis of vertical bars
Contrary to horizontal stripes, that were absent in some of the F2 individuals, none of the 230 F2 completely lacked bar patterns (Fig. 1.3). This not only suggests a rather polygenic basis, but also supports the existence of several dominant alleles that are sufficient in driving the formation of bars and ultimately override the recessive alleles that block bar formation resulting in a high robustness of the trait in the face of hybridization. As there is substantial variation in bar number, thickness and intensity, allele combinations might however result in variation of these characters. The QTL analysis showed that at least eleven different genomic loci are associated with bar number. The multiple-QTL model for the number of complete bars explained 32% of the observed variation, possibly indicating that here we were not able to detect all of the minor effect QTL due to phenotypic limitations or lack of power due to too small sample size of the QTL cross or instead indicating non-genetic effects on pattern formation. A QTL study in subsequent backcross generations instead of the F2 generation would be appropriate to better understand the effect of single QTL in a highly polygenic trait like vertical bars, as usually done when analyzing the genetic basis of highly polygenic yield-relevant traits in commercial crops (Tanksley & Nelson 1995; Bernacchi et al. 1998; Huang et al. 2003; Pillen et al. 2003). The analysis of backcross generations makes it easier to isolate certain alleles of one parental species in the genetic background that is mostly provided by the other parental species to better understand the effect of
single QTL. The fact that we were only able to detect large effect QTL for the number of complete bars but not for the total number of bars might be explained by the higher number of individuals with extreme phenotypes when only counting the complete bars (more individuals with low bar numbers).

The barred parental species of our hybrid cross (C. demasoni) usually displays either six or seven complete vertical bars (Fig. 1.1 F). In contrast, the distribution of vertical bars in the F₂ generation is shifted towards a mean value of eight bars, with some individuals showing up to 11 bars (Fig. 1.3 A). This case of transgressive segregation is another potential source for novel phenotypes resulting from hybridization and one of many examples of transgressive segregation in hybrid crosses of Lake Malawi cichlids (Parsons et al. 2011; Holzman & Hulsey 2017; Husemann et al. 2017). We speculate that the gap in the morphospace at the lower end of the bar distribution (Fig. 1.3 A) could hint at a developmental constraint during the formation of vertical bars preventing the formation of a vertical bar pattern with fewer than four bars (Smith et al. 1985). This would also explain why we see very few cichlid species with low bar numbers in the haplochromine radiations.

The genetic and cellular basis of color pattern formation

The large effect QTL on chromosome 18 encompasses the gene agrp2 (asip2b) that previously has been associated with the presence of horizontal stripes in African cichlids (Kratochwil et al. 2018). Low levels of agrp2 expression in the skin are associated with horizontal stripes in cichlids of the East African Great Lakes and knocking out the gene in formerly non-striped species resulted in the formation of horizontal stripes (Kratochwil et al. 2018). Remarkably, agrp2 expression in the skin does not differ between striped and non-striped regions in African cichlids (Liang et al. 2021). Even though the association between low levels of agrp2 and the presence of horizontal stripes seems evident, the exact mechanisms by which agrp2 expression in the skin affects the formation of stripes remain unclear.
While there is no clear evidence for the exact molecular functions of *agrp2* during horizontal stripe formation it could be speculated that the gene affects color pattern formation via the melanocortin system (Cal *et al.* 2017). The agouti gene family that *agrp2* is a member of, has been shown to act as a strong antagonist of different melanocortin receptors in mammals (Ollmann *et al.* 1997) and zebrafish (Zhang *et al.* 2010) which could affect chromatophore proliferation, pigment dispersion and pigment production (Cal *et al.* 2017). In zebrafish, *agrp2* expression is also necessary for the regulation of *pmch* and *pmchl* which are responsible for melanosome aggregation during background adaptation (Berman *et al.* 2009; Zhang *et al.* 2010). Both, *pmch* and *pmchl* were found in the intervals of different QTL detected in our study (*pmch* on chromosome 17 for DLS and MLS; *pmchl* on chromosome 7 for DLS and full bars). Further genes like *ednrb* (chromosome 18 for DLS, MLS and full bars) and *mitfa* (chromosome 5 for DLS) are not directly involved in melanin production and aggregation but affect pigment cell development (Nataf *et al.* 1996; Opdecamp *et al.* 1997). Previous studies emphasized the role of pigment cell development, migration and chromatophore interaction in the formation of horizontal stripes in zebrafish. In contrast to zebrafish, where melanophores (the melanin bearing pigment cells) are almost exclusively found in the dark regions of the color pattern (Patterson *et al.* 2014), melanophores are numerous in the regions between bars and stripes in East African cichlids. Nevertheless, the contrast between dark and light regions is reached through a combination of differential melanophore density and differential melanosome aggregation in the melanophores (Hendrick *et al.* 2019; Liang *et al.* 2020). Different melanophore densities between striped and non-striped regions could be obtained by interactions between the different chromatophore types as seen in zebrafish (Frohnhofer *et al.* 2013).

Even though our study revealed multiple genes that might be involved in color pattern formation in African cichlids, we still do not understand the exact molecular mechanisms and processes that connect these genes and the pattern formation. Further studies that build on the knowledge of color pattern formation, involved genes in zebrafish as well as in-depth investigation of pigment cell interactions will be needed for a more comprehensive understanding.
Conclusion

In this study, we found that the most common melanic patterns in East African cichlids, vertical bars and horizontal stripes are two genetically independent modules. The QTL analysis revealed a multigenic basis for both of the color patterns. Through a more detailed analysis of the stripe pattern, we could identify QTL that specifically associate with certain portions of the stripe pattern. Our study provides new insights into how genetic modularity and the evolution and combination of modifier loci might have driven the remarkable coloration diversification in the adaptive radiations of the East African Great Lake cichlids.

Acknowledgements

The authors thank Paolo Franchini, Christina Chang-Rudolf and the staff of the animal research facility of the University of Konstanz for their valuable help.
Supplementary materials

Visual phenotyping: stripe coverage
The potentially striped region is the region that is striped in the striped parental species *P. cyaneorhabdos*. For the MLS that is the distance between the most dorsal part of the operculum and the base of the caudal fin and for the DLS it is the distance between the anterior and posterior base of the caudal fin just dorsally of the upper lateral line. Depending on the posture of the fish, these distances could not always be covered in a straight line. This is why the potential MLS region was covered by three straight lines: line 1 from the most dorsal part of the operculum to the stripe-region dorsally of the anterior base of the anal fin, line 2 from the end of line 1 to the stripe-region dorsally of the posterior base of the anal fin, and line 3 from the end of line 2 to the central base of the caudal fin (Fig. S1.1). The potential DLS region was covered by two straight lines: line 1 from anterior base of the caudal fin to the point dorsally of the anterior base of the anal fin, line 2 from the end of line 1 to the posterior base of the dorsal fin. The total length of the lines equals the potentially striped region.

The proportion of the lines that runs over melanic regions equals the stripe coverage. When corrected for bars, the stripe regions that were crossed by bars were subtracted from both potentially and actually striped measurements. The procedure can be performed for all fish independent of the presence or absence of horizontal stripes. For a visualization, see supplementary Figure S1.1.
### Table S 1.1: Summary of multiple QTL models of visual measurements

<table>
<thead>
<tr>
<th>Trait</th>
<th>chromosome</th>
<th>interval min [cM]</th>
<th>position [cM]</th>
<th>interval max [cM]</th>
<th>LOD</th>
<th>PVE [%]</th>
<th>p value</th>
<th>PVE multiple QTL model [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DLS visual</strong></td>
<td>2</td>
<td>28</td>
<td>43.5</td>
<td>46.3</td>
<td>4.986</td>
<td>4.555</td>
<td>2.15E-05</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>15.1</td>
<td>33.7</td>
<td>39.7</td>
<td>3.989</td>
<td>3.608</td>
<td>0.001187</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0</td>
<td>92.5</td>
<td>92.8</td>
<td>1.899</td>
<td>1.681</td>
<td>0.016805</td>
<td></td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>0</td>
<td>5.4</td>
<td>48.2</td>
<td>4.142</td>
<td>3.752</td>
<td>0.000135</td>
<td></td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>10</td>
<td>34.6</td>
<td>25.4</td>
<td>22.283</td>
<td>24.419</td>
<td>2.05E-16</td>
<td></td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>3.8</td>
<td>9.3</td>
<td>51.4</td>
<td>1.72</td>
<td>1.52</td>
<td>0.024678</td>
<td></td>
</tr>
<tr>
<td></td>
<td>23</td>
<td>0</td>
<td>17.7</td>
<td>35.3</td>
<td>2.472</td>
<td>2.2</td>
<td>0.004908</td>
<td></td>
</tr>
<tr>
<td><strong>MLS visual</strong></td>
<td>2</td>
<td>29.1</td>
<td>39.6</td>
<td>46.1</td>
<td>2.951</td>
<td>3.534</td>
<td>0.00175</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0</td>
<td>47.4</td>
<td>64.7</td>
<td>1.4</td>
<td>1.651</td>
<td>0.04913</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>6.3</td>
<td>34.1</td>
<td>47.4</td>
<td>1.976</td>
<td>2.344</td>
<td>0.01423</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>3.5</td>
<td>4.8</td>
<td>13.3</td>
<td>2.668</td>
<td>3.186</td>
<td>0.00321</td>
<td></td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>7.1</td>
<td>28.3</td>
<td>44.9</td>
<td>1.667</td>
<td>1.971</td>
<td>0.02769</td>
<td></td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>0</td>
<td>3.9</td>
<td>50.3</td>
<td>1.693</td>
<td>2.002</td>
<td>0.02610</td>
<td></td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>18.9</td>
<td>24.5</td>
<td>28.3</td>
<td>10.826</td>
<td>14.06</td>
<td>7.64E-11</td>
<td></td>
</tr>
<tr>
<td><strong>Number of full bars</strong></td>
<td>2</td>
<td>13.6</td>
<td>16.6</td>
<td>33.5</td>
<td>4.841</td>
<td>6.833</td>
<td>2.71E-05</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0</td>
<td>14.6</td>
<td>35.2</td>
<td>3.383</td>
<td>4.293</td>
<td>0.000416</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0</td>
<td>89.3</td>
<td>92.8</td>
<td>2.193</td>
<td>3.013</td>
<td>0.008535</td>
<td></td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>0</td>
<td>3.6</td>
<td>10.9</td>
<td>2.159</td>
<td>2.966</td>
<td>0.00975</td>
<td></td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>0</td>
<td>0.2</td>
<td>39.9</td>
<td>2.824</td>
<td>3.906</td>
<td>0.002185</td>
<td></td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>0</td>
<td>23.6</td>
<td>25.7</td>
<td>2.687</td>
<td>3.711</td>
<td>0.002915</td>
<td></td>
</tr>
</tbody>
</table>

55.807

42.19192

32.86278
### Table S 1.2: Summary of multiple QTL models of patternize analysis

<table>
<thead>
<tr>
<th>Trait</th>
<th>chromosome</th>
<th>Interval min [cM]</th>
<th>Interval max [cM]</th>
<th>LOD</th>
<th>PVE [%]</th>
<th>p value</th>
<th>PVE multiple QTL model [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>DLS patternize</td>
<td>2</td>
<td>22.3</td>
<td>43.8</td>
<td>7.009</td>
<td>10.988</td>
<td>1.84E-07</td>
<td></td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>7.7</td>
<td>30.6</td>
<td>2.805</td>
<td>4.306</td>
<td>0.00177</td>
<td></td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>0</td>
<td>18.3</td>
<td>1.805</td>
<td>2.774</td>
<td>0.01615</td>
<td></td>
</tr>
<tr>
<td></td>
<td>23</td>
<td>26.6</td>
<td>33.6</td>
<td>2.272</td>
<td>3.395</td>
<td>0.00666</td>
<td></td>
</tr>
<tr>
<td>DLS patternize anterior</td>
<td>7</td>
<td>28.5</td>
<td>44.4</td>
<td>7.046</td>
<td>11.499</td>
<td>1.47E-07</td>
<td>27.06787</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>16.3</td>
<td>62.2</td>
<td>1.719</td>
<td>2.658</td>
<td>0.021526</td>
<td>24.10528</td>
</tr>
<tr>
<td></td>
<td>23</td>
<td>12.8</td>
<td>31.1</td>
<td>3.864</td>
<td>6.105</td>
<td>0.000179</td>
<td></td>
</tr>
<tr>
<td>DLS patternize posterior</td>
<td>1</td>
<td>18.5</td>
<td>61.7</td>
<td>2.955</td>
<td>3.758</td>
<td>0.00499</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0</td>
<td>43.7</td>
<td>2.907</td>
<td>4.584</td>
<td>0.00161</td>
<td></td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>9.4</td>
<td>30.4</td>
<td>4.531</td>
<td>7.264</td>
<td>4.42E-05</td>
<td></td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>0</td>
<td>18.3</td>
<td>2.629</td>
<td>4.135</td>
<td>0.00298</td>
<td></td>
</tr>
<tr>
<td>MLS patternize</td>
<td>1</td>
<td>0</td>
<td>62.7</td>
<td>1.915</td>
<td>2.807</td>
<td>0.0149</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>8.1</td>
<td>57.3</td>
<td>2.622</td>
<td>3.863</td>
<td>0.00319</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>15.6</td>
<td>56.2</td>
<td>2.215</td>
<td>3.25</td>
<td>0.00778</td>
<td></td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>0</td>
<td>46.3</td>
<td>2.191</td>
<td>3.214</td>
<td>0.0082</td>
<td></td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>19.4</td>
<td>51.4</td>
<td>6.504</td>
<td>9.968</td>
<td>6.41E-07</td>
<td></td>
</tr>
<tr>
<td>MLS patternize anterior</td>
<td>2</td>
<td>0</td>
<td>47.5</td>
<td>2.133</td>
<td>3.566</td>
<td>0.00893</td>
<td></td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>18.1</td>
<td>58.6</td>
<td>2.743</td>
<td>4.666</td>
<td>0.00231</td>
<td></td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>0</td>
<td>58.4</td>
<td>2.214</td>
<td>3.746</td>
<td>0.00746</td>
<td></td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>8.4</td>
<td>31.2</td>
<td>2.805</td>
<td>4.775</td>
<td>0.00202</td>
<td></td>
</tr>
<tr>
<td>MLS patternize posterior</td>
<td>1</td>
<td>0</td>
<td>41.5</td>
<td>2.198</td>
<td>2.784</td>
<td>0.00881</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0</td>
<td>49.5</td>
<td>1.516</td>
<td>1.907</td>
<td>0.03823</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>7.8</td>
<td>56.9</td>
<td>2.403</td>
<td>3.049</td>
<td>0.00557</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>18.8</td>
<td>56.2</td>
<td>1.877</td>
<td>2.369</td>
<td>0.0176</td>
<td></td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>0</td>
<td>55.9</td>
<td>2.769</td>
<td>3.527</td>
<td>0.00258</td>
<td></td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>18.8</td>
<td>31.1</td>
<td>8.854</td>
<td>11.998</td>
<td>5.29E-09</td>
<td></td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>11.2</td>
<td>51.4</td>
<td>2.647</td>
<td>3.367</td>
<td>0.00335</td>
<td></td>
</tr>
<tr>
<td>Bars patternize dorsal</td>
<td>2</td>
<td>1.1</td>
<td>36.8</td>
<td>2.149</td>
<td>2.799</td>
<td>0.01022</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>2.6</td>
<td>62.6</td>
<td>2.679</td>
<td>3.507</td>
<td>0.00331</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>6.7</td>
<td>49.5</td>
<td>3.18</td>
<td>4.184</td>
<td>0.00114</td>
<td></td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>0</td>
<td>63.8</td>
<td>2.201</td>
<td>2.868</td>
<td>0.00915</td>
<td></td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>0</td>
<td>58</td>
<td>2.137</td>
<td>2.782</td>
<td>0.0105</td>
<td></td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>0</td>
<td>18.8</td>
<td>6.236</td>
<td>8.465</td>
<td>1.68E-06</td>
<td></td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>0</td>
<td>26.7</td>
<td>1.639</td>
<td>2.123</td>
<td>0.03035</td>
<td></td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0</td>
<td>23.9</td>
<td>2.307</td>
<td>3.009</td>
<td>0.0073</td>
<td></td>
</tr>
<tr>
<td>Bars patternize ventral</td>
<td>7</td>
<td>33.8</td>
<td>50.7</td>
<td>9.373</td>
<td>14.919</td>
<td>9.86E-10</td>
<td></td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>20.6</td>
<td>62.9</td>
<td>2.167</td>
<td>2.364</td>
<td>0.02854</td>
<td></td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>0</td>
<td>27.1</td>
<td>3.185</td>
<td>4.759</td>
<td>0.00087</td>
<td></td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>3.9</td>
<td>51</td>
<td>2.406</td>
<td>3.567</td>
<td>0.00487</td>
<td></td>
</tr>
</tbody>
</table>
Figure S 1.1: Stripe measurements and landmarks. Each red dot shows a single landmark used for the analysis in patternize. The colored lines indicate the lines that have been used for the continuous measurement of each stripe. Lines 1 and 2 are orange and blue respectively for both DLS and MLS, line 3 for the MLS is colored in green.
Figure S 1.2 Patternize PCA for parental species, DLS and MLS. (A) shows the PCA for the DLS. (B) shows the PCA for the MLS. Red triangles represent individuals of the non-striped species *C. demasoni*, yellow circles represent individuals of the striped parental species *P. cyaneorhabdos*. The cichlid outlines on the axes show the region that was analysed and indicate the coloration that correlates with high or low values of each PC. Red coloration indicates high levels of dark coloration (stripe present), blue coloration indicated low levels of dark coloration (stripe absent).
Figure S 1.3 Patternize PCA for parental species, vertical bars. (A) shows the PCA for the dorsal part of the vertical bars. (B) shows the PCA for the ventral part of the vertical bars. Red triangles represent individuals of the non-striped species *C. demasoni*, yellow circles represent individuals of the striped parental species *P. cyaneorhabdos*. The cichlid outlines on the axes show the region that was analysed and indicate the coloration that correlates with high or low values of each PC. Red coloration indicates high levels of dark coloration (stripe present), blue coloration indicated low levels of dark coloration (stripe absent).
Figure S 1.4 Horizontal stripes and vertical bars in the F₁ generation. Individuals of the F₁ generation show horizontal stripes and vertical bars. Horizontal stripes are never fully developed in the first offspring generation.
Figure S 1.5 Effect plots of QTL for dorsolateral stripe. Dots show the mean of the phenotype, error bars indicate the standard error.
**Figure S 1.6 Effect plots of QTL for midlateral stripe.** Dots show the mean of the phenotype, error bars indicate the standard error.
Figure S 1.7 Effect plots of QTL for full bars. Dots show the mean of the phenotype, error bars indicate the standard error.
Chapter 2
Brain transcriptomics provide insights into short- and long-term regulation of aggressive behavior in the Siamese fighting fish, *Betta splendens*

Jan Gerwin and Axel Meyer

In preparation for publication

Abstract

Aggression represents one of the most common social interactions of vertebrates. A precise regulation of aggressive behavior is essential as aggression cannot only bear benefits but is often also costly. Immediate regulation of social behavior is usually controlled by the endocrine and nervous systems, but knowledge on the regulation of these systems through changes in gene expression in the brain is scarce. Fighting fish of the genus *Betta* constitute an excellent model system to study the evolution of the regulation of aggressive behavior in the brain as the different species vary in size and breeding behavior, which subsequently results in a wide range of levels of aggression. In this study, we used the highly aggressive species *Betta splendens*, a species well known for their long history of domestication and their use in competitive fights, and the closely related but less aggressive species *Betta simplex* to investigate the transcriptomic response of the brain to an aggressive interaction. Using dyadic mirror interactions, we quantified aggressive behaviors in both species. Transcriptomics of four different brain regions (diencephalon, telencephalon, cerebellum and optic lobes) were used to identify differentially expressed genes between control and treatment fish and to conduct weighted correlation network analysis (WGCNA). The behavioral essay confirmed the expected behavioral differences between the two studied species. Differential gene expression revealed expression changes of genes involved in e.g. metabolism and hormonal regulation in *B. splendens*, suggesting fast and short-term behavioral regulation. Additionally, both species showed changes in expression of
several genes involved in neuronal outgrowth, thus indicating neuronal plasticity as long-term regulation of social behavior in response to an aggressive interaction. In this study, we emphasize the importance of both short- and long-term regulation of potentially costly aggressive behaviors through transcriptional changes.
Introduction

The regulation of behavioral and physiological responses to aggressive interactions is crucial for the survival and reproduction of the individual, as aggressive conflicts with conspecifics can be costly (Haller & Wittenberger 1987; Haller 1995). While high levels of aggression can increase an individual’s fitness through the establishment of e.g. territories (Watson & Miller 1971; Deverill et al. 1999; Boerner & Kruger 2008) or the access to mating opportunities (Surbeck et al. 2012), each aggressive interaction comes with energetic costs (Haller & Wittenberger 1988; Haller 1995; Castro et al. 2006) and the risk of injuries. In some extreme cases high levels of aggression will even decrease fitness because the aggressor will neglect parental care in favor of fighting opponents (Duckworth 2006) or will be more susceptible to predation (Huntingford 1976). This fine line between benefit and cost of aggression underlines the importance of making the correct decisions in an aggressive interaction. The regulation of the behavioral response to an aggressive interaction can lead to a variety of behaviors: from a full commitment to fight to an immediate avoidance of the interaction. The decision for either response is based on information collected during the interaction (e.g. size and aggression of the opponent) but also on memorized information collected during previous social interactions (McGregor et al. 2001; Karino & Someya 2007; Dziewczynski & Leopard 2010; Dziewczynski & Perazio 2012).

In teleosts, immediate behavioral changes in response to aggressive interactions are usually controlled by the nervous system (Burmeister et al. 2005) and the endocrine system (Silva et al. 2020), but changes in gene expression of the brain can also regulate the behavior in a current but most importantly in future social interactions, by affecting the physiology and structure of the neural and endocrine systems. Two brain regions crucial for the regulation of aggressive behaviors in teleosts are the telencephalon, which in teleosts harbors regions functionally homologous to the mammalian hippocampus (Rodríguez et al. 2002; Broglio et al. 2010), and the diencephalon, harboring the hypothalamus. Both the hippocampus (Adamec ; Demski & Knigge 1971; Chang & Gean 2019) and the hypothalamus (Demski & Knigge 1971; Lin et al. 2011) are important brain regions for the regulation of aggressive behavior in vertebrates. Early studies involving the lesion of the telencephalon in the
Siamese Fighting Fish *Betta splendens* were the first to clearly demonstrate the role of the telencephalon in aggressive behaviors in teleosts (Shapiro *et al.* 1974; de Bruin 1980). The hormonal regulation of social behavior in teleosts is generally mediated by hormones of the vasopressin/oxytocin family (Oliveira & Gonçalves 2008). Activity of these hormones has been associated with aggressive and dominance behavior in both the telencephalon (Santangelo & Bass 2010; Huffman *et al.* 2012; Huffman *et al.* 2015; Lema *et al.* 2015; Rodriguez-Santiago *et al.* 2017) and the diencephalon (Greenwood *et al.* 2008; Lema 2010; Almeida & Oliveira 2015; Lema *et al.* 2015; Elkins *et al.* 2017; Eastman *et al.* 2020). High levels of gene expression changes in both brain regions have been found in the context of aggressive interactions in different teleost species, such as zebrafish (Filby *et al.* 2010), stickleback (Bukhari *et al.* 2017) and knife fish (Eastman *et al.* 2020). While changes in gene expression might not influence the immediate behavioral response to a social cue, they can influence the behavior in an ongoing social interaction and more importantly affect the behavioral responses in future social interactions by permanently adjusting both neuronal and hormonal systems (Sorensen *et al.* 2013). Aggressive interactions have been shown to permanently adjust hormone secretion (Oliveira *et al.* 2002) and influence the formation of new neurons in the brain (Sorensen *et al.* 2012) according to social experience. These permanent physiological and structural changes of the neuroendocrine system in response to environmental cues are known as neural plasticity and are crucial for the fine-tuned regulation of behavioral responses in social interactions (Sorensen *et al.* 2013).

While differences of hormonal secretion and gene expression in response to aggressive interactions have been studied extensively to explain behavioral differences within different teleost species, the role of these mechanisms in the evolution of differential behaviors between closely related species remains unclear. Species of the genus *Betta* represent a perfect model system to investigate the evolution of regulation of social behavior, and especially aggressive behavior. *B. splendens* has been a model system in ethology to study different aspects of aggressive behavior for decades (Simpson 1968; Baenninger 1984). Male-male confrontations in this species reliably produce
aggressive interactions and involve behavioral patterns that are easy to phenotype. Gill flaring, the
erection of the operculum, constitutes one of the most important behavioral patterns as it is the first
aggressive behavior demonstrated in an aggressive interaction (Simpson 1968; Forsatkar et al. 2016)
and the amount of gill flaring correlates with the general aggression level of the individual (Evans
1985). It is well established that aggressive behavior in *B. splendens* is regulated based on earlier social
experiences (McGregor *et al.* 2001; Karino & Someya 2007; Dzieweczynski & Leopard 2010;
Dzieweczynski & Perazio 2012), as well as based on current social cues (Doutrelant *et al.* 2001;
Clotfelter & Paolino 2003; Dzieweczynski & Walsh 2011; Dzieweczynski & Perazio 2012; Bertucci *et al.*
2014). Also, changes of aggression in *B. splendens* have been linked to increased levels of androgenic
hormones (Dzieweczynski *et al.* 2006; Dzieweczynski & Buckman 2013; Dzieweczynski & Hebert 2013;
Ramos *et al.* 2021) and the serotonergic system (Clotfelter *et al.* 2007; Dzieweczynski & Hebert 2012;
Eisenreich & Szalda-Petree 2015). While studies investigating the regulation of aggressive behavior
through social cues and the endocrine system are numerous, knowledge of the genetic architecture of
aggressive behavior of *B. splendens* is scarce. Only with the recent emergence of high quality genomic
information of the species (Fan *et al.* 2018; Kwon *et al.* 2022), the exploration of genetics of aggressive
behavior in *B. splendens* started (Vu *et al.* 2020). The vast number of breeding forms of *B. splendens*
with different levels of aggression (Verbeek *et al.* 2007; Kwon *et al.* 2022) and high number of closely
related species (Rüber *et al.* 2004) that also show different levels of aggression (Verbeek *et al.* 2007)
are a great system to investigate the consequences of aggressive behavior on the gene expression in
the brain of *B. splendens*.

In this study, we investigated the transcriptional response of individual brain regions of two
*Betta* species to an aggressive interaction at two different time points. We used the highly aggressive
species *B. splendens*, a species well known for and long domestication history in the context of
aggressive behaviors (Kwon *et al.* 2022) and the closely related, but less aggressive species *B. simplex*.
We created transcriptomes of the four different brain regions telencephalon, diencephalon, optic
tectum and cerebellum, after exposure to a mirror (aggression trigger) for 10 or 30 minutes. The aim
of this study is to compare the transcriptomic reaction to an aggressive interaction of the different brain regions between both species. In particular, we investigated (i) which brain regions showed the strongest transcriptional response to the aggressive interaction, (ii) what genes showed differential expression between control and treatment groups, looking both at single candidate genes and functional enrichment of all differentially expressed genes and (iii) the changes of gene co-expression networks in response to aggressive interactions to uncover the role of transcriptional changes in the short- and long-term regulation of aggressive behavior.
Materials and methods

Animal experiments and dissection
Fish of both species were housed at 27°C in 12:12 hour light-dark cycles. All fish were socially isolated for 7 days before the experiment and were transferred to a 6L experimental tank (20cm x 20cm x 15cm) (Fig. S2.1) the night before the experiment. Both of the long sides and one of the short sides of the experimental tank were permanently covered with white pieces of plastic to avoid any interaction with the environment. The remaining side of the tank was covered with a removable piece of white plastic. All experiment where filmed from the top. Each experiment started with a 15-minute baseline phase with all four sides of the aquarium covered by opaque plastic pieces. After the first phase, the removable piece of plastic was lifted and revealed depending on the treatment a 15cm x 20cm mirror that was placed directly on the outside of the tank or no mirror. The fish remained in the tank for either 10 or 30 minutes after the gate was lifted.

The fish were sacrificed immediately after the end of the experiment. The head was removed from the rest of the body and the brain of the fish was dissected in phosphate-buffered saline (PBS). The brain was divided into the four desired brain regions (Fig 2.2 C) and each part was separately stored in 1ml of RNAlater (Invitrogen). Each dissection process lasted for ca. 10 minutes. The samples were stored at RT for 60 minutes before being stored at -20°C.

Fish belonged to four different treatment groups per species: 10 minutes with exposure to the mirror, 10 minutes control (no mirror), 30 minutes with exposure to the mirror and 30 minutes control (no mirror). We used eight biological replicates per treatment group, resulting in a total of 64 individuals (32 per species).

Behavioral scoring
The top view videos were used for the behavioral scoring of the studied individuals during the treatment using the software BORIS (Friard & Gamba 2016). While males of *B. splendens* perform a variety of aggressive behaviors during an encounter with another male (bites, fin beats, lateral displays and gill flaring), we decided to focus our analysis on the usage of gill flaring, the erection of the
operculum while facing the opposing individual. Gill flaring is the most consistently used behavior (Forsatkar et al. 2016) and the scoring is highly repeatable. Even though gill flaring is only used as an aggressive display and does not harm the opposing fish, it is well established that the duration off gill flaring correlates with levels of aggression in *B. splendens* (Evans 1985). We determined the time of the start and the end of each gill flare. From the collected data, we could derive the attack latency (time between lifting the gate and the first attack), the frequency of gill flares and the total duration of gill flares during the encounter with the mirror image.

RNA extraction
For RNA isolation, the tissue was transferred into a 2ml screwcap reaction tube together with homogenization beads and 1ml of Trizol (Invitrogen) and then homogenized using a tissue lyser (Tissue Lyser II, Qiagen). The homogenized tissue was stored at RT for 5min and 1ml of homogenate was transferred into a fresh 1.5ml reaction tube. After adding 200µl of chloroform, the homogenate was mixed vigorously for 15sec. After 10 minutes of incubation at RT, the homogenate was centrifuged at maximum speed for 15 minutes at 4°C. After centrifugation 400µl of the aqueous phase (top layer) were transferred into a fresh 1.5ml reaction tube, mixed with 400µl of isopropanol and stored at RT for 15min. After incubation, the mixture was centrifuged at maximum speed for 8min at 4°C. After centrifugation, the supernatant was removed carefully and discarded, 1ml of 75% ethanol was added to wash the RNA pellet and the mixture was centrifuged at 7500g for 5min at RT. As much ethanol as possible was removed without disturbing the pellet, the tube was shortly spun down again and the rest of the ethanol was removed. The sample was now air dried until all ethanol had evaporated (3-5min). The RNA pellet was dissolved in 22µl of nuclease free water. The concentration of each sample was measured using the Qubit Fluorometer 2.0 (Invitrogen) and RNA integrity was measured via electrophoresis in a TapeStation system (Agilent).

Library preparation, sequencing and read processing
Libraries for RNA-sequencing were prepared using the SENSE mRNA-Seq Library Prep Kit (Lexxogen) and the CORALL Total RNA-Seq Library Prep Kit (Lexxogen) following the manufacturer’s instructions.
Library pools of 30 samples were generated and 150bp paired end fragments were sequenced using the Illumina HiSeq X 10 platform at the Beijing Genomics Institute (BGI). After sequencing and demultiplexing, the adapters, poly-A tails and low quality base pairs were removed using BBDDuK. Next, the STAR aligner (Dobin et al. 2013) was used to align the reads to the *B. splendens* reference genome. The resulting gene-count tables were used for differential gene expression analysis.

**Differential gene expression analysis**

For the identification of differentially expressed genes we used the DESeq2 pipeline (Love et al. 2014) in R (version 4.0.3). Using the DESeq function we compared gene expression between all 16 possible combinations of tissue (diencephalon, telencephalon, cerebellum and optic lobes), treatment (mirror and control), time (10 and 30 minutes) and species (*B. simplex* and *B. splendens*) and extracted the results for all relevant comparisons (treatment vs control of the same combination of species, tissue and time). Differentially expressed genes were chosen with a \( p \)-value cutoff of 0.05 and a false discovery rate cut off at 0.1. Gene names were identified using the getGene function of the mygene package in R. To ensure that the results were not biased due to the fact that one of the study species was more distantly related to the species of the reference genome than the other (all reads were mapped to the *B. splendens* reference genome), we repeated the alignments, but used the reference genome of the climbing perch, *Anabas testudineus*, instead. The results of the subsequent differential gene expression analysis were similar to the analysis using the *B. splendens* genome as a template, but there was an overall trend of fewer differentially expressed genes, which is why we decided to use the *B. splendens* genome as for the identification of read counts. The enrichment analysis for gene ontology was conducted using the topGO package in R and visualization of the enrichment analysis was conducted using REVIGO (Supek et al. 2011).

**Weighted correlation network analysis (WGCNA)**

To understand concerted changes of gene expression that are not necessarily reflected in the differential gene expression analysis we conducted weighted correlation network analysis using the WGCNA package (Langfelder & Horvath 2008) in R. Briefly, we compared co-expression modules of all
samples of a brain region of the same species (consensus modules) to co-expression modules that are specific to the treatment groups (mirror modules). The genes in mirror modules that did not show significant overlap with the consensus modules (mirror specific modules) were analyzed for functional enrichment. For the network analysis, we followed the WGCNA tutorial. Only genes with at least 250 counts per million reads were used to construct the co-expression networks. The minimal modules size was set to 40 genes for all comparisons. The enrichment analysis for gene ontology was conducted using the topGO package in R and visualization of the enrichment analysis was conducted using REVIGO (Supek et al. 2011).
Results

Behavioral differences between study species
Timing, number and duration of gill flares were scored for the tested individuals during the whole treatment (either 10 or 30 minutes) for both treatment groups (mirror and control) using the top view videos. Our analysis confirms our assumption of higher levels of aggression in \textit{B. splendens} when compared to \textit{B. simplex} (Fig. 2.1). Only slight differences were found for the delay of the first gill flare towards the mirror image, but there is an obvious trend of \textit{B. splendens} attacking the mirror image earlier than \textit{B. simplex} (Fig. 2.1 A). For this comparison, individuals of both time treatments were combined, as the treatment time (10 or 30 minutes) does not influence the delay of the initial attack.

Individuals of \textit{B. splendens} also spend more time gill flaring than individuals of \textit{B. simplex}. Most \textit{B. splendens} individuals of the 10 minute treatment spent more than between 20\% and 70\% of the time gill flaring, while most individuals of \textit{B. simplex} spend between 20\% and 25\% of the time gill flaring. The differences are less pronounced for the 30 minute treatment (Fig. 2.1 B). Furthermore, individuals of \textit{B. splendens} show a higher frequency of gill flares compared to individuals of \textit{B. simplex} (Fig. 2.1 C).

Almost all individuals of \textit{B. splendens} show the behavior between 3 and 11 times per minute, while almost all individuals of \textit{B. simplex} show the behavior at a lower frequency than 2.5 times per minute in the 10 minute treatment. This comparison is the only one showing significant differences between the two study species (Wilcoxon signed-rank test, $p = 0.0089$). Again, the differences for the 30 minute treatment are less pronounced. Frequency and duration of gill flaring are generally lower in the 30 minute treatment, because gill flaring is replaced by other displays and aggressive behaviors in later stages of an aggressive interaction (data not shown).

Transcriptomes of brain regions
After quality filtering the similarity of the remaining 252 transcriptomic samples of the two species (\textit{B. splendens} and \textit{B. simplex}) and four brain regions (telencephalon, diencephalon, optic lobes, cerebellum) was analyzed in multidimensional scaling plots. The analysis revealed that samples group according to tissue and species (Fig. 2.2). Samples of the same species were more similar to each other.
than samples of the same tissue of different species. For example, samples of the telencephalon of \textit{B. splendens} were most similar to samples of the diencephalon of \textit{B. splendens}, rather than to the telencephalon of \textit{B. simplex} (Fig. 2.2 A). Within species, the relative position of samples within the MDS plot is the same for both species: samples of diencephalon and telencephalon are similar in \textit{B. splendens} and \textit{B. simplex}, the same is true for samples of the optic lobes and the cerebellum. These results indicate a general transcriptional differentiation between the species resulting in the clear split between species on dimension 1, but also a conservation of transcriptional differences between brain regions, as the relative position of brain region within each species is almost the same (Fig. 2.2 A).

\textbf{Figure 2.1 Behavioral differences between the study species \textit{B. simplex} and \textit{B. splendens}.} The panels show the delay until the first gill flaring during after the beginning of the treatment (A), the time spend gill flaring during the treatment (B) and the frequency of gill flares during the treatment (C) comparing the two study species for each time and mirror treatment. s: seconds; min: minutes; *: \(p \leq 0.01\); NS: not significant.
A more detailed look at the samples of each species confirms the similarity of the different tissues within species. A clear differentiation between control and mirror samples could not be detected in the MDS plots, indicating that the genes driving the general transcriptional differences between tissues are not the same genes responding to the behavioral treatment (Fig. 2.2 B, C).

Transcriptional response to aggressive interaction
To find genes related to the regulation of aggressive behavior we looked for genes that show differential expression between individuals that were exposed to their mirror image compared to the control individuals. Our analysis revealed differentially expressed genes in all brain regions for both time treatments in *B. simplex* and *B. splendens*. While each brain region of *B. simplex* showed a transcriptional reaction to the aggressive interaction we found clear differences between the different brain regions regarding the number of differentially expressed genes (Fig. 2.3 A). The diencephalon of *B. simplex* showed to weakest transcriptional reaction to the treatment: only seven genes were up-regulated in the 10 minute treatment and only one gene was down-regulated in the 30 minute treatment. The telencephalon and optic lobes showed a stronger response for the 30 minute treatment with 27 upregulated and 43 downregulated genes, respectively (Fig. 2.3 A). In the cerebellum of *B. simplex* the strongest transcriptional response was found for the 10 minute treatment with 29 downregulated genes (Fig. 2.3 A). There was no clear pattern of differential gene expression among time treatments or brain regions for *B. simplex*.

When compared to the results from *B. simplex*, the transcriptional response to the aggressive interaction in *B. splendens* was stronger, reflected by the higher number of differentially expressed genes (Fig. 2.3 B). In contrast to *B. simplex*, the diencephalon of *B. splendens* showed the strongest response of the four brain regions. 121 and 169 genes were up and down regulated in the 10 minute treatment group, respectively, and 67 and 110 genes were up and downregulated in the 30 minute treatment. The number of differentially expressed gene was similar but slightly lower in the telencephalon of *B. splendens*. The transcriptional response of the cerebellum and optic lobes of *B. splendens* was weaker when compared to the telencephalon and diencephalon with less than 10 up or
down-regulated genes for each brain region and treatment group, except the 27 down-regulated genes in the optic lobes for the 10 minute treatment. All four brain regions show a stronger transcriptional response for the 10 minute treatment than for the 30 minute treatment.

Figure 2.2 Multidimensional scaling of transcriptome samples. The panels show the similarity of all transcriptomes in a 3D MDS plot (A) and the samples of *B. simplex* (C) and *B. splendens* (D) in 2D MDS plots. Gray areas in C and D highlight regions occupied by samples of the same tissue. Dark and light dots depict control (light) and mirror (dark) samples within each tissue. B shows a sketch of a sagittal cut through the brain of the two studied species. Dim: dimension, FC: Fold change of expression.

A comparison of the genes that were up and down regulated for the different time treatments revealed that there was little overlap of differentially expressed genes between the 10 and 30 minute treatments. While there was no overlap of up or down-regulated genes between the two time points in none of the four brain regions in *B. simplex* (Fig. S2.2), we detected a small overlap between time treatments in all brain regions of *B. splendens*, but never more than five genes per comparison (Fig.
These results indicate rapid shifts in gene expression changes during an aggressive encounter and unique transcriptional responses at the different time points in both species.

**Figure 2.3 Up- and down-regulation of genes in different brain regions.** The bar plots show the number of up and down regulated genes in different brain regions of *B. simplex* (A) and *B. splendens* (B) at different treatment time points. Dark bars show down-regulated, light bars show up-regulated genes.

**Candidate genes and pathways**

For a better understanding of the gene function that showed differential expression between treatment and control samples, we conducted a GOterm enrichment analysis for the differentially expressed genes for each brain region and time point. The similarity of enriched GOterms was visualized in revigo-like MDS plots (Fig. 2.4 and Fig. 2.5). Due to the lower number of differentially expressed genes, *B. simplex* shows fewer enriched functional terms than *B. splendens*. While there are 18 enriched functional terms for the telencephalon of *B. simplex* after ten minutes, none of the closely related terms seems to play a role in the regulation of aggression (Fig. 2.4 A). There was only one enriched term for the 30 minute treatment of the diencephalon of *B. simplex* (Fig. 2.5 C). Eight of the enriched functional terms in the diencephalon of *B. splendens* after 10 minutes are related to the regulation of carbohydrate metabolism, indicating changing energy demands in this brain region due to the aggressive interaction (Fig. 2.5 B). The analysis of the diencephalon of *B. splendens* after 30
minutes revealed the highest number of groups of closely related functional terms. Differentially expressed genes had functions in hormonal regulation and sex determination, both indicating changes in the hormonal system, and histone modification and DNA replication, indicating changes in gene expression and cell proliferation (Fig. 2.5 D). The analysis of the telencephalon revealed a lower number of enriched functional terms with similar functions. *B. simplex* showed differential expression of genes involved in tissue differentiation after 30 minutes (Fig. 2.4 C). For *B. splendens*, we found genes associated with axogenesis and serotonin biosynthesis after 10 minutes, and genes related to DNA replication after 30 minutes (Fig. 2.4 B).

Figure 2.4 Functional enrichment in differentially expressed genes of the telencephalon. The revigo like MDS graphs depict similarity of functionally enriched terms in the differentially expressed genes in the telencephalon of *B. simplex* after 10 minutes (A) and 30 minutes (C) and of *B. splendens* after 10 minutes (B) and 30 minutes...
(D). Each dot represents one functionally enriched term. Grey clouds indicate groups of closely related functional terms. Only functional terms relevant to the regulation of aggression are labelled.

Figure 2.5 Functional enrichment in differentially expressed genes of the diencephalon. The revigo like MDS graphs depict similarity of functionally enriched terms in the differentially expressed genes in the diencephalon of *B. simplex* after 10 minutes (A) and 30 minutes (C) and of *B. splendens* after 10 minutes (B) and 30 minutes (D). Each dot represents one functionally enriched term. Grey clouds indicate groups of closely related functional terms. Only functional terms relevant to the regulation of aggression are labelled.

These results suggest that the differentially expressed genes in both brain regions of *B. simplex* are not restricted to functions that would be expected to play a role in the regulation of aggressive behavior. Genes differentially expressed in *B. splendens* have functions in metabolism, hormone regulation, histone modification and DNA replication, that can all play a role in the short and long term regulation of aggressive behaviors. The diencephalon seems to be more adaptable than the
telencephalon suggested by the high number of enriched functional terms associated with the regulation of aggressive behavior in this brain region.

While most genes that show differential expression between treatments do not seem related to aggressive behavior, an analysis of the top six differentially expressed genes of every comparison revealed genes in the diencephalon and telencephalon of both species that might play an active role in the response to the aggressive interaction (Tab. 2.1, Tab. 2.2 and Tab. 2.3). One group of genes that show notable changes in both brain regions and species are genes of the myosin family. Expression of the myosin light chain (\textit{mylpfa}) and the myosin heavy chain are upregulated in the diencephalon of \textit{B. splendens} after 10 minutes, while only the expression of the myosin light chain is upregulated in the diencephalon of \textit{B. simplex} after 10 minutes. In addition, the myosin heavy chain is downregulated in the telencephalon of \textit{B. splendens} after 30 minutes. The expression of another myosin gene, \textit{myosin7} (\textit{myh7}), is downregulated in the diencephalon of \textit{B. splendens} after 30 minutes, in the telencephalon of \textit{B. splendens} after 10 minutes and in the telencephalon of \textit{B. splendens} after 30 minutes.
### Table 2.1 Top six differentially expressed genes in B. simplex

The table presents the top six up- and down-regulated genes in the telencephalon and diencephalon of *B. simplex*. In some cases, fewer than six or no genes were differentially expressed. The column "ensembl gene id" shows only the last five digits of the gene ID preceded by “ENSBSLG000000”.

<table>
<thead>
<tr>
<th>Species</th>
<th>Brain region</th>
<th>Time treatment</th>
<th>direction</th>
<th>ensembl gene id</th>
<th>gene name</th>
<th>log2Fold Change</th>
<th>p adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>B. simplex</td>
<td>Telencephalon</td>
<td>10</td>
<td>up</td>
<td>11069</td>
<td>NA</td>
<td>17.9642107</td>
<td>1.16E-11</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>11628</td>
<td>NA</td>
<td>2.17896189</td>
<td>0.061813</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>down</td>
<td>19639</td>
<td>NA</td>
<td>-17.7651847</td>
<td>8.31E-22</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>00065</td>
<td>polyunsaturated fatty acid lipoxigenase ALOX15B-like</td>
<td>-17.1839996</td>
<td>5.25E-12</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4790</td>
<td>pleckstrin homology domain containing, family F (with FYVE domain) member 1</td>
<td>-16.2027363</td>
<td>9.27E-25</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>05061</td>
<td>Metazoan signal recognition particle RNA</td>
<td>-15.4547592</td>
<td>6.50E-17</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>03162</td>
<td>centromere protein P</td>
<td>-4.14623247</td>
<td>1.12E-05</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>19152</td>
<td>probable E3 ubiquitin-protein ligase HERC3</td>
<td>-2.69038162</td>
<td>5.56E-06</td>
</tr>
<tr>
<td></td>
<td></td>
<td>30</td>
<td>up</td>
<td>04790</td>
<td>pleckstrin homology domain containing, family F (with FYVE domain) member 1</td>
<td>16.4436096</td>
<td>2.18E-19</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>14051</td>
<td>NA</td>
<td>5.71374258</td>
<td>0.037134</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>18764</td>
<td>iroquois homeobox 4a</td>
<td>5.09995782</td>
<td>0.00772</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3960</td>
<td>NA</td>
<td>4.8999787</td>
<td>0.031622</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>15868</td>
<td>platelet endothelial cell adhesion molecule-like</td>
<td>4.55632966</td>
<td>0.047838</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>25909</td>
<td>T-box transcription factor 16</td>
<td>4.37172388</td>
<td>0.079605</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>down</td>
<td>09700</td>
<td>NA</td>
<td>-16.040631</td>
<td>6.04E-12</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>22268</td>
<td>parvalbumin alpha</td>
<td>-14.0805083</td>
<td>2.40E-15</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>15027</td>
<td>SMAD family member 9</td>
<td>-2.92027874</td>
<td>0.012885</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>12025</td>
<td>uncharacterized LOC114849351</td>
<td>-2.35336289</td>
<td>0.071926</td>
</tr>
<tr>
<td></td>
<td>Diencephalon</td>
<td>10</td>
<td>up</td>
<td>11069</td>
<td>NA</td>
<td>21.8691501</td>
<td>1.09E-19</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>20561</td>
<td>myosin heavy chain, fast skeletal muscle-like</td>
<td>17.9135221</td>
<td>1.09E-19</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10425</td>
<td>NA</td>
<td>11.5404733</td>
<td>0.033068</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>17875</td>
<td>tumor necrosis factor b (TNF superfamily, member 2)</td>
<td>4.65272685</td>
<td>0.083752</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10624</td>
<td>myostatin b</td>
<td>3.11374912</td>
<td>0.009543</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>24098</td>
<td>annexin A1a</td>
<td>2.707949397</td>
<td>0.083752</td>
</tr>
<tr>
<td></td>
<td></td>
<td>30</td>
<td>down</td>
<td>21475</td>
<td>phospholipase C, eta 2b</td>
<td>-3.8216083</td>
<td>0.096073</td>
</tr>
</tbody>
</table>
Table 2.2 Top six differentially expressed genes the diencephalon of B. splendens. The table presents the top six up- and down-regulated genes in the diencephalon of *B. splendens*. The column “ensembl gene id” shows only the last five digits of the gene ID preceded by “ENSBSLG000000”.

<table>
<thead>
<tr>
<th>Species brain region</th>
<th>Time treatment</th>
<th>direction</th>
<th>ensembl gen id</th>
<th>gene name</th>
<th>log2Fold Change</th>
<th>padj</th>
</tr>
</thead>
<tbody>
<tr>
<td>B. splendens diencephalon</td>
<td>10</td>
<td>up</td>
<td>4295</td>
<td>myosin light chain, phosphorylatable, fast skeletal muscle a</td>
<td>8.24796685</td>
<td>5.08E-05</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>20561</td>
<td>myosin heavy chain, fast skeletal muscle-like</td>
<td>7.25139058</td>
<td>0.009536</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>22268</td>
<td>parvalbumin alpha</td>
<td>6.79175794</td>
<td>0.001039</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>04850</td>
<td>tropomin T, fast skeletal muscle isoforms-like</td>
<td>6.2319128</td>
<td>0.006363</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>02538</td>
<td>phosphorylase, glycogen, muscle A</td>
<td>5.96994907</td>
<td>1.33E-05</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>13142</td>
<td>leukotriene B4 receptor 1-like</td>
<td>5.59089558</td>
<td>0.00224</td>
</tr>
<tr>
<td></td>
<td></td>
<td>down</td>
<td>02048</td>
<td>NA</td>
<td>-22.23038</td>
<td>2.42E-40</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>23225</td>
<td>retinal pigment epithelium-derived rhodopsin homolog</td>
<td>-5.91165308</td>
<td>0.002872</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>22923</td>
<td>NA</td>
<td>-5.8886601</td>
<td>0.000213</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>16273</td>
<td>collagenase 3-like</td>
<td>-5.86044704</td>
<td>0.091906</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>10644</td>
<td>homeobox C6a</td>
<td>-5.42400029</td>
<td>0.017868</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>14971</td>
<td>TPA-induced transmembrane protein</td>
<td>-5.29344242</td>
<td>0.002872</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>up</td>
<td>01858</td>
<td>NA</td>
<td>21.2508384</td>
<td>1.74E-14</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>11829</td>
<td>bone gamma-carboxyglutamate (gla) protein, like</td>
<td>18.1080886</td>
<td>7.20E-28</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>05414</td>
<td>chemokine (C-X-C motif) ligand 19</td>
<td>17.4715541</td>
<td>5.38E-13</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>17310</td>
<td>FA complementation group C</td>
<td>5.29889777</td>
<td>0.002993</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>13142</td>
<td>leukotriene B4 receptor 1-like</td>
<td>5.13195926</td>
<td>0.015514</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>22268</td>
<td>parvalbumin alpha</td>
<td>5.03882258</td>
<td>0.044364</td>
</tr>
<tr>
<td></td>
<td></td>
<td>down</td>
<td>09162</td>
<td>myosin-7</td>
<td>-21.564742</td>
<td>2.30E-30</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>02696</td>
<td>NK6 homeobox 3</td>
<td>-5.52559129</td>
<td>0.001198</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>22604</td>
<td>solute carrier family 39 member 13</td>
<td>-5.46125375</td>
<td>5.02E-06</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>22832</td>
<td>protein disulfide-isomerase TMX3-like</td>
<td>-4.98212884</td>
<td>0.001089</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>19176</td>
<td>probable E3 ubiquitin-protein ligase HERC4</td>
<td>-4.86656226</td>
<td>0.015514</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>16327</td>
<td>solute carrier family 35 member F2-like</td>
<td>-4.84323575</td>
<td>0.045979</td>
</tr>
</tbody>
</table>
Table 2.3 Top six differentially expressed genes the telencephalon of *B. splendens*. The table presents the top six up- and down-regulated genes in the telencephalon of *B. splendens*. The column “ensembl gene id” shows only the last five digits of the gene ID preceded by “ENSBSLG000000”.

<table>
<thead>
<tr>
<th>Species</th>
<th>brain region</th>
<th>Time</th>
<th>direction</th>
<th>ensembl gene id</th>
<th>gene name</th>
<th>log2Fold Change</th>
<th>padj</th>
</tr>
</thead>
<tbody>
<tr>
<td>B. splendens</td>
<td>telencephalon</td>
<td>10</td>
<td>up</td>
<td>1858</td>
<td>NA</td>
<td>20.7808727</td>
<td>2.07E-13</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13194</td>
<td>four and a half LIM domains 2a</td>
<td>17.6559151</td>
<td>1.84E-32</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>01982</td>
<td>matrix metalloproteinase-15-like</td>
<td>5.67159028</td>
<td>0.019348</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10420</td>
<td>ST6 (alpha-N-acetylneuraminyl-2,3-beta-galactosyl-1,3)-N-acetylneuraminic acid-2,6-sialyltransferase 2</td>
<td>5.66380709</td>
<td>0.022666</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>17612</td>
<td>interleukin-21 receptor</td>
<td>5.51580997</td>
<td>8.99E-08</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>15079</td>
<td>solute carrier family 24 member 1</td>
<td>5.50753206</td>
<td>4.14E-08</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>down</td>
<td>24156</td>
<td>NA</td>
<td>-23.62392</td>
<td>1.14E-07</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>24525</td>
<td>NA</td>
<td>-21.6435479</td>
<td>9.90E-34</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>17287</td>
<td>NA</td>
<td>-16.4246771</td>
<td>2.56E-11</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>19384</td>
<td>phosphoglycerate mutase 2 (muscle)</td>
<td>-15.63474</td>
<td>7.95E-19</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>09162</td>
<td>myosin-7</td>
<td>-15.6270876</td>
<td>4.41E-13</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>02096</td>
<td>ras-related protein Rap-2a-like</td>
<td>-5.38902027</td>
<td>0.000221</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>30</td>
<td>23087</td>
<td>homeobox A2b</td>
<td>27.53225</td>
<td>2.96E-69</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>24156</td>
<td>NA</td>
<td>24.5062961</td>
<td>1.27E-08</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1858</td>
<td>NA</td>
<td>19.1582574</td>
<td>2.68E-11</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>11829</td>
<td>bone gamma-carboxyglutamate (gla) protein, like</td>
<td>17.6059791</td>
<td>1.22E-25</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13194</td>
<td>four and a half LIM domains 2a</td>
<td>16.550269</td>
<td>6.31E-26</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4174</td>
<td>troponin I type 2a (skeletal, fast), tandem duplicate 2</td>
<td>15.2725962</td>
<td>1.35E-13</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>22296</td>
<td>NA</td>
<td>-17.5059399</td>
<td>8.61E-19</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9162</td>
<td>myosin-7</td>
<td>-16.6658831</td>
<td>1.89E-15</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>19384</td>
<td>phosphoglycerate mutase 2 (muscle)</td>
<td>-14.731115</td>
<td>3.12E-15</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>20561</td>
<td>myosin heavy chain, fast skeletal muscle-like</td>
<td>-14.5942806</td>
<td>3.01E-11</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2941</td>
<td>cyclin-O</td>
<td>-4.94353462</td>
<td>0.041949</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>12962</td>
<td>solute carrier family 51 subunit alpha</td>
<td>-4.93856602</td>
<td>0.077325</td>
</tr>
</tbody>
</table>
Myosins have key functions in the development of new neurons, axonal transport and synaptic functions (Brown & Bridgman 2004) and are hence promising candidate genes that might play a role in the neuronal response to aggressive social interactions. Another promising gene that showed differential expression between control and treatment individuals is parvalbumin alpha (pvalb). This gene associated with aggressive behaviors in mice (van Heukelum et al. 2019) was upregulated in the diencephalon of B. splendens after 10 and 30 minutes and showed decreased expression in the telencephalon of B. simplex after 30 minutes.

---

**Figure 2.6 Functional enrichment in genes of the mirror-specific modules of the diencephalon.** The revigo like MDS graphs depict similarity of functionally enriched terms in the genes contained in the mirror-specific modules in the diencephalon of B. simplex after 10 minutes (A) and 30 minutes (C) and of B. splendens after 10 minutes (B) and 30 minutes (D). Each dot represents one functionally enriched term. Grey clouds indicate groups of closely related functional terms. Only functional terms relevant to the regulation of aggression are labelled.
Co-expression gene networks

In addition to identifying differentially expressed genes, we investigated the concerted changes of gene expression between treatment and control by identifying co-expression gene networks that are unique to the treatment group and exploring the genes of these networks for functional enrichment related to neuronal functions and neurogenesis. The number of new co-expression modules in each treatment group ranged from three (diencephalon *B. splendens* after 30 minutes) (Fig. S2.7) to 15 (telencephalon *B. splendens* after 30 minutes) (Fig. S2.6) but we could not observe a general trend indicating a higher number of new modules in a certain species or tissue (Fig. S2.4, Fig. S2.5, Fig. S2.6 and Fig. S2.7).
Figure 2.7 Functional enrichment in genes of the mirror-specific modules of the telencephalon. The revigo like MDS graphs depict similarity of functionally enriched terms in the genes contained in the mirror-specific modules in the telencephalon of *B. simplex* after 10 minutes (A) and 30 minutes (C) and of *B. splendens* after 10 minutes (B) and 30 minutes (D). Each dot represents one functionally enriched term. Grey clouds indicate groups of closely related functional terms. Only functional terms relevant to the regulation of aggression are labelled.

The analysis of gene function in the mirror specific co-expression modules revealed an enrichment of gene function in neuronal processes and neurogenesis in the telencephalon of both species, especially for the 30 minute treatment in *B. simplex* (Fig. 2.6) and the 10 minute treatment in *B. splendens* (Fig. 2.7). Genes in the new modules in the diencephalon groups rarely had neuronal functions in either species. This trend cannot be attributed to a higher number of new modules or higher number of genes in the new modules in the telencephalon compared to the diencephalon and
instead indicates changes of expression of genes related to neurogenesis and neuronal functions in the telencephalon.
Discussion

Our study aimed at understanding the transcriptional of different brain regions in response to an aggressive interaction in the two fighting fish species, \textit{B. splendens} and \textit{B. simplex}. Our main goal was to identify transcriptional differences associated with the different levels of aggression observed between the two study species. We found differential expression and changes in co-expression networks involving genes with functions in short- and long-term regulation of aggressive behavior.

Rapid transcriptional response in the brain of \textit{B. splendens}

Our analysis of differential gene expression in the four brain regions (optic lobes, cerebellum, diencephalon and telencephalon) revealed clear differences in the transcriptional response to an aggressive interaction between the highly aggressive species \textit{B. splendens} and the more peaceful species \textit{B. simplex}. Overall, the number of differentially expressed genes is higher in \textit{B. splendens} than in \textit{B. simplex} indicating a stronger transcriptional response to the social interaction. A closer look at the number of genes in the different brain regions and time treatment shows, that \textit{B. splendens} shows the strongest response in the telencephalon and the diencephalon. These results are in concordance with transcriptomic analysis associated with aggression and dominance behavior in zebrafish, \textit{Danio rerio}, that showed the strongest transcriptional response in the same brain regions (Filby \textit{et al.} 2010).

All brain regions of \textit{B. splendens} showed a stronger transcriptional response for the 10-minute treatment. In \textit{B. simplex} on the other hand, the strongest transcriptional response was found in the optic lobes and no consistent differences between the time treatments could be found. These results are in conflict with previous results from transcriptomic analysis of \textit{B. splendens}, where more differentially expressed genes were found after 60 minutes when compared to a 20-minute treatment and almost all differentially expressed genes where upregulated in the treatment group (Vu \textit{et al.} 2020). It is likely that these differences arise due to the use of whole brain tissue in Vu \textit{et al.} (2020) concealing brain region specific transcriptional differences. Further, studies in the three-spined stickleback (\textit{Gasterosteus aculeatus}) showed similar results to our study: the number of up or down regulated genes is similar, but again the strongest transcriptional response can be found after 60
minutes (Bukhari et al. 2017). Both aforementioned studies analyzed brain transcriptomes at different treatment time points but the earliest time point was set after 20 minutes and 30 minutes, respectively, probably not expecting strong transcriptional changes after a shorter period of time. Our results show that only ten minutes after the beginning of a social interaction, a strong transcriptional response can be observed, indicating that changes of gene expression in the brain might not only have long term effects but can also be used to regulate an ongoing social interaction, as fights between males of *B. splendens* can last considerably longer than 10 minutes (Simpson 1968; Evans 1985; Bronstein 1994; Vu et al. 2020). Evidence for rapid transcriptional changes in the context of aggressive social interactions is rare but has been found in cichlid fishes (Almeida et al. 2019) and honey bees (Harrison et al. 2019). The rapid transcriptional changes in the brain of *B. splendens* might be adaptive in an environment with frequent aggressive interactions, as fast regulation of behavioral responses and metabolism are crucial to avoid the potentially high costs of an aggressive interaction. The lack of concerted expression changes in the brain of *B. simplex* after both 10 and 30 minutes indicates a less important role of behavioral regulation in this species, which could be rooted in different breeding behaviors, leading to varying fitness benefits from aggressive behaviors. Contrary to *B. splendens*, the mouth breeding males of *B. simplex* do not rely on the establishment of a territory for the construction of a bubble nest to ensure reproduction.

Short-term regulation of social behavior through changes in metabolism and endocrine system

For a comprehensive overview and a better functional understanding of differentially expressed genes, we analyzed the sets of differentially expressed genes in the telencephalon and the diencephalon for each treatment group, looking also for enriched functional terms. The top six up- or down-regulated genes in these brain regions were further investigated. In particular, we looked for the enrichment of genes with functions in neuronal processes, neurogenesis, metabolism and DNA replication. Genes with functions in the regulation of carbohydrate metabolism are significantly enriched among the differentially expressed in the diencephalon of *B. splendens* after 10 minutes. The increased activity of the brain during a social interaction might result in an increased energy demand, which can result in
the switch of the primary energy source in the cell. Accordingly, changes in carbohydrate metabolism have been linked to aggressive behavior in honeybees and *Drosophila melanogaster* (Li-Byarlay *et al.* 2014; Chandrasekaran *et al.* 2015).

After 30 minutes, differentially expressed genes with functions in hormone regulation, hormone synthesis and sex determination were significantly enriched in the diencephalon of *B. splendens*. The role of hormones of the vasopressin/oxytocin family in the regulation of aggressive behavior has been described for a number of teleost species (Greenwood *et al.* 2008; Lema 2010; Santangelo & Bass 2010; Huffman *et al.* 2012; Almeida & Oliveira 2015; Huffman *et al.* 2015; Lema *et al.* 2015; Elkins *et al.* 2017; Rodriguez-Santiago *et al.* 2017; Eastman *et al.* 2020). Furthermore, the effects of androgenic hormones, that usually play a role in sex determination, on aggressive behavior have been described for *B. splendens*. It could be shown that aggressive interactions lead to increased levels of 11-ketotestosterone and testosterone (Ramos *et al.* 2021) and that audience type and nesting status influence both aggression and 11-ketotestosterone levels in *B. splendens* (Dzeweczynski *et al.* 2005; Dzeweczynski *et al.* 2006), demonstrating the important role of androgens in the regulation of aggressive behaviors in this species.

One of the most noticeable results regarding the analysis of differentially expressed genes are the expression changes of different genes of the myosin gene family in the diencephalon and the telencephalon of both *B. splendens* and *B. simplex*, with some genes showing increased and other genes showing decreased expression after the aggressive interaction. While myosins play a central role in muscle contraction, they are no obvious candidate genes to regulate aggressive behaviors. However, it is well established that myosins play important roles in axonal transport (Bridgman 2004), and therefore increased levels of myosin could be an immediate response to the increased demands in axonal transport, as more neurotransmitter and receptors are needed at the synapses during the social interaction.

The analysis of differentially expressed genes revealed changes in the endocrine system and carbohydrate metabolism of *B. splendens* in addition to potential changes in the neuronal system of
both species *B. simplex* and *B. splendens*, all of which might have immediate effects on the regulation of the ongoing aggressive interaction. The fact that changes of the endocrine system and metabolism were exclusive to *B. splendens* suggests a more fine-tuned short-term regulation of aggressive behaviors in *B. splendens*. This might be a potential adaptation to the high level of aggression found in this species, as precise regulation of social behavior is crucial in an environment of constant aggressive encounters.

Neuronal plasticity and long-term regulation of social behavior

In addition to genes that have the potential to regulate aggressive behaviors immediately, our analysis also revealed parts of the transcriptional response that indicate permanent, long-term regulation of aggressive behavior. In particular, we were interested in genes that indicate the formation of new neurons and new connections between existing neurons. The formation of new neurons and neuronal connections in the adult individual is a process called neuronal plasticity and is crucial for the regulation of behavior and learning in the adults (Sweatt 2016).

The functional enrichment analysis of the differentially expressed genes in the diencephalon and the telencephalon show a functional role in cell proliferation in both species. Both the telencephalon and the diencephalon of *B. simplex* showed differential expression of genes involved in tissue differentiation and development. In *B. splendens*, genes involved in DNA replication were differentially expressed in both brain regions, suggesting increased rates of mitosis. Furthermore, the telencephalon of *B. splendens* revealed differential expression of genes with functions in axogenesis and serotonin biosynthesis, further indicating physiological changes in the neuronal system.

Next to their function in axonal transport, genes of the myosin family play a crucial role in the formation of novel neurons. Axons of newly emerging neurons navigate through the existing tissue depending on a structure called the growth cone. *Myosin II* plays a crucial role in the movement of the growth cone (Brown & Bridgman 2003; Brown & Bridgman 2004) and increased myosin expression in the tissue could hint at increased growth of axons of existing and new neurons. Increased expression of the myosin heavy chain was observed in the diencephalon of both species, suggesting increased
neuronal outgrowth in both species, which subsequently suggests the new growth of axons from new or existing neurons in the brain.

Another gene with the potential to have long-term effects on the regulation of social behavior is parvalbumin alpha (pvalb) that was upregulated in the diencephalon of B. splendens after 10 and 30 minutes and showed decreased expression in the telencephalon of B. simplex after 30 minutes. The expression of pvalb has been associated with dementia (Bernstein et al. 2011) and the ability of vocal learning in songbirds (Hara et al. 2012). Additionally, the differential distribution of pvalb-expressing neurons in the brains of mice has been linked different levels of aggressive behavior (van Heukelum et al. 2019) indicating that pvalb might play a role in the long term regulation of social behavior.

Conclusion

The analysis of transcriptomes in the context of an aggressive interaction revealed clear differences in the intensity of the transcriptomic response between the highly aggressive species B. splendens and the closely related, but less aggressive species B. simplex. Our analysis revealed concerted transcriptomic changes in B. splendens involving the differential expression of genes with functions in carbohydrate metabolism in the brain and hormonal regulation that might play a role in the short-term regulation of aggressive behavior. Furthermore, both species showed changes in expression level and co-expression of genes with functions in neurogenesis suggesting the long-term regulation of behavior through neuronal plasticity. Our results emphasize the need for fine-tuned regulation of potentially costly behaviors and highlight the important metabolic and neuronal adaptations to increased levels of aggression in B. splendens.

Acknowledgements

The authors thank Nadiya Orel, Sina Rometsch, Paolo Franchini, Nidal Karagic and the staff of the animal research facility of the University of Konstanz for their valuable help.
**Figure S 2.1 Experimental setup.** The panels show a schematic representation (A) and a photograph (B) of the experimental setup used for the mirror exposure treatment. The green line indicates the position of the gate that is lifted after the acclimation time. The blue line indicates the position of the mirror in the treatment samples.
Figure S 2.2 Venn diagrams showing overlap of up and downregulated genes in B. simplex. The venn diagrams show the number of shared differentially expressed genes between time treatments for diencephalon (A), telencephalon (B), optic lobes (C) and cerebellum (D) in B. simplex.
Figure S 2.3 Venn diagrams showing overlap of up and downregulated genes in B. splendens. The venn diagrams show the number of shared differentially expressed genes between time treatments for diencephalon (A), telencephalon (B), optic lobes (C) and cerebellum (D) in B. splendens.
Figure S 2.4 Heat map comparing consensus and mirror modules in the telencephalon of B. simplex. The panels show the similarity of consensus modules (x-axis) and mirror modules (y-axis) in the telencephalon of B. simplex after 10 minutes (A) and 30 minutes (B). Numbers in the table indicate number of genes shared between consensus and mirror module. Significance of overlap is depicted by red coloration, the darker the higher the significance (log($p$-value)). Mirror specific modules are labeled in green. Numbers on the axis indicate the number of genes in the corresponding module. Grey modules contain genes that could not be assigned to a module.
Figure S 2.5 Heat map comparing consensus and mirror modules in the diencephalon of B. simplex. The panels show the similarity of consensus modules (x-axis) and mirror modules (y-axis) in the diencephalon of B. simplex after 10 minutes (A) and 30 minutes (B). Numbers in the table indicate number of genes shared between consensus and mirror module. Significance of overlap is depicted by red coloration, the darker the higher the significance (log(p-value)). Mirror specific modules are labeled in green. Numbers on the axis indicate the number of genes in the corresponding module. Grey modules contain genes that could not be assigned to a module.
Figure S 2.6 Heat map comparing consensus and mirror modules in the telencephalon of _B. splendens_. The panels show the similarity of consensus modules (x-axis) and mirror modules (y-axis) in the telencephalon of _B. splendens_ after 10 minutes (A) and 30 minutes (B). Numbers in the table indicate number of genes shared between consensus and mirror module. Significance of overlap is depicted by red coloration, the darker the higher the significance (log(p-value)). Mirror specific modules are labeled in green. Numbers on the axis indicate the number of genes in the corresponding module. Grey modules contain genes that could not be assigned to a module.

### Table 2.6.1: Heat map of consensus and mirror modules in the telencephalon of _B. splendens_.

**A. Telencephalon, 10 minutes**

<table>
<thead>
<tr>
<th>Module</th>
<th>Consensus</th>
<th>Mirror</th>
<th>Consensus</th>
<th>Mirror</th>
<th>Consensus</th>
<th>Mirror</th>
<th>Consensus</th>
<th>Mirror</th>
<th>Consensus</th>
<th>Mirror</th>
</tr>
</thead>
<tbody>
<tr>
<td>mir turquoise: 1316</td>
<td>1</td>
<td>147</td>
<td>641</td>
<td>123</td>
<td>33</td>
<td>16</td>
<td>114</td>
<td>8</td>
<td>99</td>
<td>21</td>
</tr>
<tr>
<td>mir green: 149</td>
<td>0</td>
<td>5</td>
<td>43</td>
<td>6</td>
<td>12</td>
<td>0</td>
<td>11</td>
<td>2</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>mir purple: 86</td>
<td>0</td>
<td>1</td>
<td>31</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>4</td>
<td>13</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>mir red: 129</td>
<td>7</td>
<td>0</td>
<td>74</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>14</td>
<td>19</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>mir black: 119</td>
<td>4</td>
<td>17</td>
<td>41</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>31</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>mir pink: 99</td>
<td>0</td>
<td>1</td>
<td>7</td>
<td>1</td>
<td>7</td>
<td>0</td>
<td>21</td>
<td>5</td>
<td>4</td>
<td>52</td>
</tr>
<tr>
<td>mir magenta: 88</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>43</td>
<td>0</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>mir brown: 471</td>
<td>49</td>
<td>5</td>
<td>24</td>
<td>42</td>
<td>130</td>
<td>1</td>
<td>110</td>
<td>22</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>mir yellow: 252</td>
<td>9</td>
<td>0</td>
<td>5</td>
<td>3</td>
<td>104</td>
<td>0</td>
<td>62</td>
<td>11</td>
<td>1</td>
<td>57</td>
</tr>
<tr>
<td>mir tan: 43</td>
<td>3</td>
<td>0</td>
<td>17</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>mir blue: 955</td>
<td>20</td>
<td>17</td>
<td>156</td>
<td>0</td>
<td>9</td>
<td>37</td>
<td>597</td>
<td>25</td>
<td>0</td>
<td>45</td>
</tr>
<tr>
<td>mir greenyellow: 58</td>
<td>0</td>
<td>6</td>
<td>7</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>19</td>
<td>0</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>mir grey: 9</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
</tr>
</tbody>
</table>

**B. Telencephalon, 30 minutes**

<table>
<thead>
<tr>
<th>Module</th>
<th>Consensus</th>
<th>Mirror</th>
<th>Consensus</th>
<th>Mirror</th>
<th>Consensus</th>
<th>Mirror</th>
<th>Consensus</th>
<th>Mirror</th>
<th>Consensus</th>
<th>Mirror</th>
</tr>
</thead>
<tbody>
<tr>
<td>mir magenta: 79</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>mir purple: 73</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>mir blue: 1056</td>
<td>18</td>
<td>31</td>
<td>52</td>
<td>22</td>
<td>43</td>
<td>51</td>
<td>1410930</td>
<td>28</td>
<td>564</td>
<td>60</td>
</tr>
<tr>
<td>mir yellow: 109</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>mir brown: 140</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>mir salmon: 66</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>mir lightyellow: 41</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>mir green: 105</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>10</td>
<td>7</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>mir midnightblue: 52</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>mir black: 97</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>mir turquoise: 1321</td>
<td>46</td>
<td>40</td>
<td>10786</td>
<td>6711046</td>
<td>2574</td>
<td>8714</td>
<td>246530</td>
<td>285046</td>
<td>62</td>
<td>81</td>
</tr>
<tr>
<td>mir grey: 60:43</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>8</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>mir lightgreen: 41</td>
<td>0</td>
<td>2</td>
<td>7</td>
<td>2</td>
<td>1</td>
<td>6</td>
<td>0</td>
<td>13</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>mir tan: 69</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>mir red: 101</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>mir greenyellow: 71</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>21</td>
</tr>
<tr>
<td>mir lightcyan: 51</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>mir pink: 91</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>16</td>
<td>0</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>mir cyan: 54</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>mir grey: 12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

_Cons: Consensus, Mir: Mirror_
Figure S 2.7 Heat map comparing consensus and mirror modules in the diencephalon of B. splendens. The panels show the similarity of consensus modules (x-axis) and mirror modules (y-axis) in the diencephalon of B. splendens after 10 minutes (A) and 30 minutes (B). Numbers in the table indicate number of genes shared between consensus and mirror module. Significance of overlap is depicted by red coloration, the darker the higher the significance (log(p-value)). Mirror specific modules are labeled in green. Numbers on the axis indicate the number of genes in the corresponding module. Grey modules contain genes that could not be assigned to a module.
Chapter 3
Seeing red: evolution of cone opsin genes in the Siamese fighting fish, *Betta splendens*

Jan Gerwin, Axel Meyer and Julián Torres-Dowdall

In preparation for publication

Abstract

Gene duplication is one of the most important substrates for the accumulation of novel genotypic diversity and the subsequent evolution of phenotypic diversity. Understanding the evolutionary history of duplicated genes and the resulting functional changes of the gene product is crucial for a comprehensive understanding of adaptive evolution. A family of genes that commonly experience duplication events with subsequent sub- and neofunctionalization are the visual opsin genes of teleost fishes. While duplication of the green-sensitive opsins rh2 is common in teleosts, only few cases of multiple duplication events of the red-sensitive opsin lws are known. Factors restricting or driving the expansion of lws genes in the teleost genomes remain unknown. In this study, we investigate the visual system of the anabantoid species *Betta splendens* with a focus on the five paralogs of lws found in this species. Using whole genome sequences of nine anabantoid species, we investigate the evolutionary history of lws genes in this order. Our results show that at least two independent duplications of lws occurred in the Betta lineage. The analysis of amino acid sequences of the lws paralogs of Betta revealed high levels of diversification in four of the seven transmembrane regions of the lws protein. Amino acid substitutions at two key-tuning sites lead to differentiation of absorption maxima ($\lambda_{max}$) between the paralogs within Betta. Finally, eye transcriptomics of *B. splendens* at different developmental stages revealed expression shifts between paralogs for all cone opsin classes. The lws genes are expressed according to their relative position in the opsin cluster throughout ontogeny. We
conclude that the temporal collinearity of \textit{lws} expression might facilitate sub-and neofunctionalization of \textit{lws} in \textit{Betta} and teleost opsins in general.

Introduction

Exploring the origin of genetic diversity that eventually can result in phenotypic diversity is one of the main goals of evolutionary biology. Even though the emergence of novel genetic variants is a prerequisite for selection, retaining novel genotypic diversity in protein coding sequences can often be problematic due to pleiotropic effects, even if the new resulting phenotype is potentially desirable (Molodtsova \textit{et al.} 2014). One way to circumvent the problem of pleiotropic effects can be the rise of new genetic material by gene duplication (Ohno 1970). In this case, the sequence and function of one paralog of the gene can be maintained, while the second copy can accumulate mutations that then can be object to selection possibly resulting in a new function of this paralog (Rastogi & Liberles 2005).

In most cases, gene duplication is the result of either a tandem duplication of a small part of the genome or a whole genome duplication. The stem lineage of teleost fishes experienced a fish specific genome duplication (Meyer & Van de Peer 2005) that resulted in the expansion of many gene families that we can still see today when comparing teleost and tetrapod genomes (Amores \textit{et al.} 1998; Cortesi \textit{et al.} 2021) and has been discussed to be a major factor in the success of teleost divergence (Hoegg \textit{et al.} 2004; Glasauer & Neuhauss 2014). Duplication in tandem, that results in two paralogs in close proximity to each other, has been a major driver of diversity in some teleost gene families, too (Peatman & Liu 2007; Lu \textit{et al.} 2012; Rennison \textit{et al.} 2012). However, the most common fate of the new paralog after duplication is a loss of function, as mutations that impair the gene’s original function are likely to accumulate after the release of selective pressure (Lynch & Conery 2000). Another possible fate for the duplicates is subfunctionalization (de Souza \textit{et al.} 2005; Kluver \textit{et al.} 2005). In this case, both copies take over part of the original function of the gene, e.g. expression during certain ontogenetic stages. This allows for the accumulation of genetic divergence among copies increasing their functional divergence and fine-tuning adaptation (Wagner 1998; Lynch & Force 2000; Flagel &
A great example of a gene family that commonly experience duplication events with subsequent subfunctionalization are the opsin genes of teleost fishes. The visual system of teleost fishes is highly diverse, as different species have adapted to a wide range of light conditions of different aquatic environments (Carleton et al. 2020). The typical teleost genome harbors two classes of rod opsin genes. One of these genes, the extra-ocular rhodopsin (exorh), is not expressed in the retina and has no function in vision (Mano et al. 1999; Bellingham et al. 2003). The second class (rh1) is expressed in the retina and responsible for vision in dim light. Furthermore, there are typically four classes of cone opsin genes responsible for color vision at different wavelengths (sws1, sws2, rh2, lws) (Bowmaker 1998; Yokoyama 2000; Musilova et al. 2019). The number of genes in each of the classes can vary substantially as duplication events happened frequently during the evolutionary history of opsin genes (Rennison et al. 2012; Musilova et al. 2019). While a single copy of rh1 is present in most teleost taxa, few deep-sea lineages expanded the number of rh1 paralogs drastically in response to low levels of visible light in their environment (Musilova et al. 2019). Cone opsin genes show a more dynamic evolutionary history driven by ancient and more recent duplication events and subsequent gene losses, resulting in a plethora of different opsin repertoires across the teleost phylogeny (Bowmaker 2008; Cortesi et al. 2015; Cortesi et al. 2021; Musilova & Cortesi 2021). The UV and blue-light sensitive genes sws1 and sws2 experienced the lowest degree of duplication of the cone opsin genes, with most species possessing one copy of sws1 and one to three copies of sws2 (Cortesi et al. 2015). One to three copies of the green-sensitive opsin rh2 are present in most teleost genomes (Musilova et al. 2019) with expansions to up to eight copies of the gene in some marine species (Musilova et al. 2019; Musilova & Cortesi 2021). One or two paralogs of the red-sensitive opsin lws can be found in most teleost genomes, but this gene has experienced recent duplication events in many different lineages resulting in 4 or 5 paralogs in the genomes of some species (Cortesi et al. 2021). Reconstructing the evolutionary history of cone opsin genes remains challenging due to gene conversion among paralogs (Rennison et al. 2012; Musilova et al. 2021).
The high number of paralogs within different opsin classes is putatively maintained by molecular divergence resulting in variation in spectral sensitive of the different copies allowing for fine-tuned changes during ontogeny or in response to environmental changes (Chinen et al. 2003; Ward et al. 2008; Torres-Dowdall et al. 2017; Carleton et al. 2020). The effects of amino acid substitutions at certain positions in the opsin protein have been studied extensively allowing to deduct changes in the spectral sensitivity for some substitutions (Yokoyama 1995; Yokoyama 2008; Yokoyama et al. 2008b; Yokoyama & Jia 2020). The changes in spectral sensitivity after an amino acid substitution can be small (Yokoyama et al. 2008a), but the accumulation of multiple substitutions at key positions might result in major spectral sensitivity changes (Watson et al. 2011), which can help to compensate for evolutionary losses of other opsin classes (Dulai et al. 1999; You et al. 2014). Substitutions of the opsin’s residues that interact with the light absorbing chromophore play a critical role in causing spectral shifts of the opsin (Yokoyama 2008). While amino acid changes drove the diversity of opsin genes on a large phylogenetic scale resulting in the wide range of different opsin genes we find in teleosts today, sequence changes have been used to adapt to new photic conditions on short evolutionary time scales as well (Register et al. 1994; Terai et al. 2001; Ward et al. 2008; Rennison et al. 2012; You et al. 2014; Harer et al. 2018; Cortesi et al. 2021).

Even though most species only express a subset of opsins at the same time, the high diversity of opsin genes allows different teleost species to adapt their visual sensitivity to a wide range of photic environments and to changing needs throughout their lifetime. Accordingly, most teleost species express different sets of opsins throughout their ontogeny, most typically shifting from shortwave sensitive genes during early ontogenetic stages to a set of more longwave sensitive genes during the adult live stage (Spady et al. 2006; Carleton et al. 2008; Shand et al. 2008; Allison et al. 2010; Harer et al. 2017; Chang et al. 2021). Furthermore, the presence of several opsin genes with different spectral sensitivities allows for quick plastic changes of the expressed opsin profile in response to changes in the photic environment (Hofmann et al. 2010; Harer et al. 2017; Nandamuri et al. 2017; Sakai et al. 2018; Harer et al. 2019). The possibility to change the spectrum of visual sensitivity by changing
expression levels of different opsin genes also allows fast adaptation to new photic environments. Studies from various young lineages that inhabit new photic environments indicate that the fastest way to adapt to new photic conditions is the acquisition of fixed changes of expression of the existing opsin repertoire (Carleton & Kocher 2001; Parry et al. 2005; Carleton et al. 2010; O’Quin et al. 2010; Rennison et al. 2016; Torres-Dowdall et al. 2017; Harer et al. 2018; Wright et al. 2019; Torres-Dowdall et al. 2021). These adaptive changes in the visual system of teleost are only made possible by the combination of frequent duplication events followed subfunctionalization of the paralogs by amino acid substitutions.

Here, we investigate the visual system of the Siamese Fighting Fish, *B. splendens*. *B. splendens* has been domesticated hundreds of years ago (Kwon et al. 2022) and a multitude of different color and fin morphs have been bred through artificial selection (Wang et al. 2021). The species has been a popular model organism in behavioral biology for decades (Baenninger 1966; Simpson 1968; Bronstein 1994), but with the recent emergence of high quality genomic information on the species, genetic studies become more abundant (Wang et al. 2021; Zhang et al. 2021; Kwon et al. 2022). In particular, we explore the evolutionary history of the opsin genes in Anabantiformes and examine how gene duplication has affected the evolution of amino acid sequence and expression of the different opsin genes with an emphasis on the long wavelength sensitive opsin (*lws*) as recent studies have shown that the genome of *B. splendens* harbors the unusually high number of five copies of that gene (Cortesi et al. 2021). Specifically, we investigated (i) the evolutionary history of opsin genes in the order Anabantiformes to determine when opsin gene duplications occurred, (ii) the role of subfunctionalization by looking at sequence divergence that could result in changes in spectral sensitivity and (iii) the expression pattern of opsins in *B. splendens* during ontogeny using long-read RNA sequencing, which allows for precise mapping of reads to the highly similar paralogs of the *lws* gene.
Materials and methods

Sequence collection

Opsin sequences were extracted from whole genomes of ten different species: Danio rerio, Amphilophus citrinellus, Channa argus, Anabas testudineus, Helostoma temminckii, B. splendens, B. imbellis, B. mahachaiensis, B. smaragdina and B. siamorientalis. Sequences of D. rerio, Am. citrinellus, A. testudineus and B. splendens were extracted via the search function of the ensemble genome browser (www.ensembl.org). Whole genome sequences of B. smaragdina and B. siamorientalis were available as short reads. The short reads were aligned to the reference genome of B. splendens using the Burrows-Wheeler Aligner (Li & Durbin 2009) and sorted and indexed using samtools. Consensus sequences from the alignments were created using the mpileup command of samtools, bcftools and vcfutils.pl. Whole genome assemblies of B. imbellis, B. mahachaiensis and H. temminckii were available. Sequences of species that were not available on ensembl were extracted from the reference genomes using ViroBLAST (Deng et al. 2007) by blasting the coding sequence of the first exon of the Betta splendens orthologue of each gene. The Integrative Genomics Viewer (Robinson et al. 2011) was used to extract the desired region from each genome. Sequences of each gene were aligned in seaview (Gouy et al. 2010) using the muscle alignment algorithm and alignments were refined by hand afterwards. After alignment, intronic regions were removed for all further analysis if not stated differently.

Fish husbandry and tissue collection

A single brood of Betta splendens offspring was separated from the parents and was incubated in a large petri dish at 28°C in a 12:12 dark:light cycle. Three days post fertilization, the eggs hatched and the free-swimming larvae were moved to a 500ml plastic container. Larvae were fed infusoria starting at five days post fertilization and freshly hatched artemia starting seven days post fertilization. The fish were moved to aerated 2L plastic containers at seven days post fertilization. At 28 days post fertilization, the fish were moved to 16L tanks in a flow through water system. Fish were fed frozen
adult artemia and red mosquito larvae starting about 8 weeks after fertilization.

For tissue collection, fish were euthanized using tricaine mesylate (MS222). Individuals at 3dph, 7dph, 16dph, 20dph, 30dph, and 40dph and younger were transferred to RNAlater immediately after they were euthanized, stored at room temperature for one hour and then stored at 4°C until further processing. Juvenile's (120dph) and adult's (270dph) eyes were dissected, transferred to RNAlater (Invitrogen), stored at room temperature for one hour and then stored at 4°C until further processing. Individuals from 40dph and younger were full siblings, juveniles and adults were derived from the same breeding stock.

RNA isolation and Nanopore sequencing

To determine the developmental pattern of opsin gene expression in *B. splendens*, RNA was isolated from whole eye tissue from individuals at each sampled developmental stage. As isolation from eyes of individuals from the first two developmental stages (3dph and 7dph) did not yield enough RNA for sequencing, we pooled two heads per developmental stage for an increased yield during RNA isolation. RNA was extracted using TRIzol (ambion©) reagents following the standard trizol-chloroform protocol (Rio et al. 2010). The isolated RNA was prepared for Nanopore long read sequencing using the PCR-cDNA barcoding kit (SQK-PCB109, Oxford Nanopore Technologies) following the manufacturer’s instructions. Briefly, we used total (non-poly-A selected) RNA for reverse transcription followed by PCR (17 amplification cycles) and barcoding of the samples. After pooling, the 12 samples were sequenced using a single MinION Flow Cell (R9, FLO-MIN106D, Oxford Nanopore Technologies). We chose a long-read sequencing technique over short-read sequencing, as the later might result in elevated mapping errors due to the high similarity among the different *lws* paralogs. An example of this is the annotation of the *B. splendens* genome in the *ensembl* genome browser, where several transcripts span the different *lws* paralogs.
Read processing and expression analysis

After demultiplexing, we used Pychopper (v.2.5.0) for identification, orientation, trimming and filtering (-Q 10) of full-length Nanopore reads. Reads shorter than 1000bp were discarded as these were frequently mapped to multiple locations in the genome. We used the nanopore-res-isoforms pipeline to align the remaining reads to the *B. splendens* reference genome. Alignments were visually inspected and the number of reads aligned to each opsin gene were determined using the Integrative Genomics Viewer (Robinson et al. 2011). Proportional expression of each cone opsin gene was calculated as relative expression of all cone genes at a given developmental stage following (Harer et al. 2018).

Estimating species relationships

The species phylogeny was created using Orthofinder (Emms & Kelly 2019). Orthofinder uses proteomes to infer species relationships. Coding sequences were collected from ensembl (*B. splendens, D. rerio, Am. citrinellus* and *A. testudineus*) or created by aligning short reads to the reference genome of *B. splendens* and then extracting coding regions using *gffread* and the *B. splendens* annotation (*H. temminckii, B. imbellis, B. mahachaiensis, B. smaragdina* and *B. siamorientalis*). The phylogeny was estimated using multiple sequence alignments (-M msa) instead of inferring it from the orthogroups.

Estimating evolutionary history of opsin genes

Gene trees estimating the relationships of the different opsin genes were generated using IQ-TREE 2 (Minh et al. 2020). Aligned coding sequences were used for the automated model selection (Kalyaanamoorthy et al. 2017) and 1000 bootstraps were performed for each gene tree.

We searched for potential breakpoints in the opsin gene alignments indicative of recombination or gene conversion using the Genetic Algorithm for Recombination Detection (GARD) (Kosakovsky Pond et al. 2006) on datamonkey (www.datamonkey.org). The analysis was conducted at the gene class level with the following settings: Run mode: faster; Genetic Code: Universal code; Site-to-site rate variation: General discrete, Rate classes: 4. The synteny of the intergenic regions of opsin
genes was analyzed using R. Sequence were compared in a sliding windows of 25bp using a step size of 3bp. Regions that are more than 65% identical appear as a data point in the synteny dot plot.

Amino acid similarity
Similarity in the amino acid sequences of the different opsin gene paralogs within each spectral class was explored using sliding windows analyses. For the within Betta comparison, every paralog of each gene of each species was analyzed separately. Each position of the amino acid sequence of the protein was compared to the homolog position of all other paralogs of LWS1e resulting in a number indicating the percentage of sequences with an identical amino acid. The higher the value the higher similarity with the LWS1e paralogs will be. The mean between the paralogs of all species was calculated for each paralog and then displayed in a sliding window of 30 amino acids in the final plot. The comparison between Betta and other teleost species was performed in the same way, now comparing the Betta sequences to the sequence of LWS proteins of the other teleost species.
Results

Opsin genes in Anabantiformes
In all the anabantiform species analyzed, we identified one copy of the extra-ocular rhodopsin (exorh) and one copy of the retinal rhodopsin (rh1, Fig. 3.1 A). The cone opsin gene repertoire was found to be more variable (Fig. 3.1 A). All anabantiform species have lost the UV sensitive opsin sws1 and show two paralogs of the blue sensitive opsin sws2. Two copies of the green sensitive opsin rh2 were identified in the genomes of C. argus and H. temminckii, but three paralogs were found in A. testudineus and all the Betta species. The highest variation in paralog number was found for the red sensitive gene lws: while in the non-Betta species only two lws copies were identified, the genomes of all Betta species contain five paralogs of this opsin gene. We did not find copy number variation for any opsin gene among the different Betta species (Fig. 3.1 A).

Evolutionary history of lws
The lws gene tree mostly reflects the anabantiform species tree: lws paralogs within species are more closely related to each other than they are to potential orthologs form other species. The exception to this is within the genus Betta where orthologs lws genes cluster together (Fig. 3.1 B). These results suggest independent lws duplication events in the lineages leading to C. argus, A. testudineus, H. temminckii and Betta. Alternatively, gene conversion among paralogs within each of these lineages might have resulted in more similar lws sequences within than between species. Using GARD, we found evidence of a single recombination breakpoint at position 730 in the lws alignment, suggesting at least one potential gene conversion event. However, separate phylogenetic analyses of the regions up- and downstream of the breakpoint did not result in different relationship among lws genes of the different species. Both gene trees again largely reflected the species phylogeny (Fig S3.1).
Figure 3.1 Evolutionary history of opsin genes in the Anabantiformes. A shows a cladogram depicting the relationship of the anabantiform species and outgroups used in this study with illustration of opsin genes found in the genome of each taxon. Species of the genus Betta are depicted as a single taxon as there is no evidence of opsin gene number variation within the genus. Physical linkage among opsin genes is depicted using connecting lines. The orientation of the opsin genes in relation to neighboring genes is shown by the shape of the illustrated genes in a 5’ - 3’ direction. A phylogenetic reconstruction of the coding sequences of all lws paralogs suggest that lws gene expansion occurred in a common ancestor of the genus Betta (B).

To further explore if the two lws paralogs found in each of the more basal anabantiform species (C. argus, H. temminckii and A. testudineus) originated through independent duplication events, we analyzed the intergenic regions in a syntenic dot plot. Pairwise comparisons between all three species
and *B. splendens* revealed regions of high similarity in the region between *lws1a* and *lws1b* in all pairwise comparisons. In some cases, parts of the intergenic region are more similar than the protein coding sequences (Fig 3.2, Fig S3.3). These results indicate a common ancestry of the regions and hence suggest a single duplication event in the common ancestor of Anabantiformes that resulted in the presence of at least two *lws* paralogs in all anabantiform species (Fig 3.2).

**Evolutionary history of *rh2* and *sws2***

Similar to *lws*, the analysis of the coding sequences of *rh2* resulted in a gene tree mainly reflecting the species phylogeny. The different *rh2* paralogs cluster according to species or genus, again suggesting independent duplication events in the different lineages or high levels of gene conversion (Fig S3.2 A). The analysis with GARD revealed two breakpoints indicating recombination events between the different paralogs. As for *lws*, phylogenies based on the single gene segments between breakpoints did not result in a phylogeny reflecting the history of duplications of the paralogs (Fig S3.4).

The analysis of the *sws2* sequences resulted in a phylogeny that does not reflect the species tree but the evolutionary history of the genes. *sws2a* and *sws2b* cluster into two distinct groups and within each of those groups, the genes reflect the species phylogeny. This indicates that the presence of two *sws2* paralogs in the Anabantiformes is the result of an ancient gene duplication in a common ancestor of cichlids and Anabantiformes. (Fig S3.2 B)
Figure 3.2 Synteny dot plot of genomic region harbouring \textit{lws1a} and \textit{lws1b}. The panels present synteny of the genomic region harboring \textit{lws} paralogs \textit{lws1a} and \textit{lws1b} in \textit{A. testudineus} and \textit{H. temminckii} (A) and \textit{A. testudineus} and \textit{B. splendens} (B). Each black dot high sequence similarity in the genomes of the two compared species. Grey boxes indicate positions exons of the two \textit{lws} paralogs.
Table 3.1 Estimated maximal absorption of lws genes in B. splendens. The table shows the estimated maximal absorption of lws genes in B. splendens and the ancestral lws sequence based on amino acid substitutions at sites of the five key tuning sites of the “five-sites” rule (Yokoyama et al. 2008b).

<table>
<thead>
<tr>
<th>Tuning site</th>
<th>Ancestral</th>
<th>Betta splendens</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>lws</td>
<td>lws1a lws1b lws1c lws1d lws1e</td>
</tr>
<tr>
<td>Bo. taurus</td>
<td>164</td>
<td>S A A S S</td>
</tr>
<tr>
<td>B. splendens</td>
<td>177</td>
<td>H H H H H</td>
</tr>
<tr>
<td></td>
<td>181</td>
<td>H H H H H</td>
</tr>
<tr>
<td></td>
<td>261</td>
<td>Y Y Y Y Y</td>
</tr>
<tr>
<td></td>
<td>269</td>
<td>T T T T T</td>
</tr>
<tr>
<td></td>
<td>292</td>
<td>A A A A A</td>
</tr>
</tbody>
</table>

Estimated $\lambda_{\text{max}}$ (nm) 560 545 555 560 560 560

Amino acid substitutions and spectral sensitivity

Amino acid substitutions between paralogs of the same opsin class are crucial for changes in spectral sensitivity and subsequent subfunctionalization of different paralogs. A comparison of the amino acid sequences of all LWS paralogs in all five Betta species used in this study showed that 26.4% of all amino acid residues are variable between paralogs and species. When excluding the LWS1a paralogs from the analysis only 15% of the amino acid residues are variable, indicating a higher similarity among the remaining four paralogs. For a better understanding of the location of differentiation of amino acid sequences in Betta, we conducted a sliding window analysis comparing the amino acid sequences of opsin of the different Betta species to a representative set of opsin sequences of other teleost species. The same analysis was used to compare the amino acid sequences of opsin paralogs within the genus Betta, to find regions that have undergone differentiation during the process of potential subfunctionalization.
Figure 3.3 Amino acid diversity of LWS between Betta and other teleosts and within Betta. Sliding window analysis of Betta LWS amino acid sequence divergence. Each red line represents the mean divergence between one LWS paralog of all Betta species and LWS sequences of a set of representative teleost species (A) or the mean divergence between one LWS paralog of all Betta species and the remaining four paralogs in Betta (B). Grey areas indicate the position of transmembrane domains of LWS. Dark grey bars indicate the positions of LWS key tuning sites (based on Yokoyama et al. (2008b)).

High rates of amino acid substitutions between Betta and other teleosts were found in four different regions that correspond to four of the seven transmembrane domains of the opsin protein. The only exception is LWS1a that shows relatively high amino acid similarity at transmembrane domains III and VI. Overall, LWS1a is the paralog that is most similar to other teleost lws genes. The
remaining four paralogs show equally low levels of amino acid similarity (Figure 3.3 A). The comparison of the amino acid sequences of the LWS paralogs within the genus Betta shows that LWS1a shows the lowest level of similarity with LWS1e. LWS1d shows the highest similarity with LWS1e, LWS1b and LWS1c show intermediate values, reflecting the evolutionary relationship of the different LWS paralogs. Again, differentiation between LWS1a and the remaining paralogs is highest in the same four different regions that correspond to four of the seven transmembrane domains. Two of these regions of differentiation are in close proximity to three of the five key tuning sides that have been detected to be important in spectral tuning of lws genes (Figure 3.3 B). The sliding window analysis of the amino acid sequences of RH2 revealed, that there are three regions of differentiation between the paralogs of the genus Betta and the rh2 genes of other teleost species. This time, only two of those regions correspond to transmembrane regions of the genes and one of them is located at one of the extracellular domains of the protein. None of the three paralogs shows higher differentiation than the other two paralogs (Fig S3.5 A). The within comparison of Betta paralogs showed only one region of strong differentiation corresponding to one of the areas found in the within teleost comparison. The loss of the other two regions of differentiation in the within Betta comparison indicates, that these differentiation in these two regions is exclusive to Betta and not just a signal due to inter-paralog differences (Fig S3.5 B). The within teleost comparison of SWS2 revealed two regions of differentiation, one of them located at the second extracellular domain, the second one located at the fourth transmembrane of the protein. The within Betta comparison shows a similar pattern of differentiation between the two paralogs in Betta, indicating the differences found in the teleost wide comparison just reflect the general differences between the two paralogs that can be found in most teleost species (Fig S3.6).

Our analysis revealed differences at the amino acid level in the paralogs of LWS and RH2 that cannot be attributed to the ancient differentiation of the different paralogs. While one of the lws paralogs is more similar to the rest of teleost lws genes than the other four paralogs, all three copies of rh2 have diverged from the general teleost rh2. Differences in the two paralogs of sws2 seem to
reflect the differences found between paralogs of that class of opsins in all teleost species.

While it is not possible to deduce the spectral sensitivity of each of the opsin paralogs from the amino acid sequence, amino acid substitutions at certain key sites of each opsin class allow us to estimate relative changes of the maximum spectral sensitivity between different paralogs of the same opsin class. Two of the five key sites to tune spectral sensitivity of lws show variation between paralogs in all analyzed Betta species. A shift from alanine (A) to serine (S) at residue 177 (residue 160 in bovine rhodopsin) (A177S) that happened in LWS1c, LWS1d and LWS1e when compared to the other two paralogs is expected to increase the maximal sensitivity, as well as the substitution F274Y (261) that can be found in all paralogs but LWS1. These two amino acid substitutions indicate a low $\lambda_{\text{max}}$ for LWS1, an intermediate $\lambda_{\text{max}}$ for LWS1b and the highest $\lambda_{\text{max}}$ for the remaining three paralogs (Table 3.1). The substitution E130Q (122) at the only key tuning site in RH2b indicates an increased $\lambda_{\text{max}}$ of RH2b, too. The three key tuning sites of the sws2 gene did not show any variation between paralogs or species.

Cone opsin expression during ontogeny of Betta splendens

With proportional expression levels between 0.4 and 0.6, lws1a is the dominant opsin until 40dph, both in comparison to the other lws paralogs and genes of the other opsin classes. Expression of lws1a decreases in juvenile and adult individuals but does not fully cease. The second paralog lws1b starts being expressed at low levels from 7dph on, considerably increasing expression at 40dph and replacing lws1a as the most highly expressed gene in juveniles and adults. lws1c is expressed at low levels from 3dph on and shows increased expression in adults only. The last two paralogs generally show low expression levels. Expression of lws1d could be detected at all stages except 3dph and 30dph and is highest in adults. lws1e showed low expression levels only in juveniles and adults. Except for lws1, all lws paralogs show a trend of increased expression towards later live stages. Adults express all five paralogs of lws at the same time (Figure 3.4 C).

A similar pattern of expression can be found for the green wavelength sensitive genes. rh2a shows high expression levels at 3dph and 7dph and then decreases steadily towards adulthood. Starting at 16dph, the expression levels of rh2b increase constantly, replacing rh2a as the rh2 paralog
with the highest expression at 40dph and further increasing its relative expression in juveniles and adults. *rh2c* show almost undetectable levels of expression during some stages (16dph, 30dph, 90dph) and low levels of expression in adults. As already seen in *lws*, adults express all different paralogs of *rh2* at the same time (Figure 3.4 B).

Figure 3.4 Ontogenetic change of relative expression of different opsin paralogs. Bars in panels A-C represent the relative expression of each paralog of *sws2* (A), *rh2a* (B) and *lws1* (C) at one of the eight investigated ontogenetic stages. Error bars at stages 120 and 270 indicate standard deviation. Stacked bars in D show the sum of relative expression of all paralogs of each opsin class within ontogenetic stages.

The expression patterns of the blue sensitive opsins differ from pattern found for *lws* and *rh2*. Both paralogs, *sws2a* and *sws2b* are expressed starting at 3dph. While expression levels of *sws2a* increase until 16dph and stay constant until adulthood, expression of *sws2b* decreases though time and stops at 30dph. Contrary to *lws* and *rh2*, both paralogs of *sws2* are expressed at the same time during early ontogenetic stages and only one of them is expressed during adulthood (Figure 3.4 A).

Our analysis shows, that all of the 10 cone opsin genes found in the genome of *B. splendens* are expressed during at least one ontogenetic stage. Expression repertoires differ between early and later developmental stages. Generally, expression of few dominant paralogs of each opsin class in the
early developmental stages is replaced by expression of a more diverse set of multiple paralogs during later developmental stages, resulting in the expression of nine different cone opsin genes at the same time in adult individuals. At all developmental stages, \textit{lws} constitutes more than half of total opsin gene expression. \textit{rh2} expression is highest during the first two stages, while relative expression of \textit{sws2} peaks from 16dph to 40dph (Figure 3.4 D).
Discussion

Evolution of cone opsin gene number in Anabantiformes

Our search for opsin genes in the genomes of the study species revealed both losses and duplications of different opsin genes in the Anabantiformes. All species lost sws1, suggesting that this occurred in the common ancestor of Anabantiformes. The complete loss of sws1 genes (Lin et al. 2017; Musilova et al. 2019) or the reduction or loss of sws1 expression in adults (Hofmann et al. 2009; Escobar-Camacho et al. 2017; Musilova & Cortesi 2021; Torres-Dowdall et al. 2021) is common in teleost fishes occurring in a variety of aquatic habitats and has been described in detail in multiple teleost groups (Liu et al. 2019; Escobar-Camacho et al. 2020; Hauser et al. 2021). Oxidative stress after absorption of UV light in the retina can damage the eye (Ivanov et al. 2018), which might have led to the evolution of an UV-absorbing lens in Anabantiformes, as shown for the species *H. temminckii* and *Trichopodus leeri* (Douglas & McGuigan 1989) leading to the redundancy of sws1. Eventual losses of spectral sensitivity in the non-UV low wavelengths through the loss of sws1 could be compensated by changes in spectral sensitivity of sws2 paralogs, which are expressed in early stages as shown in this study.

We did not find variation across species in the number of sws2 paralogs, as all of the investigated anabantiform species possess two copies of that gene. This is in line with previous findings suggesting that sws2 underwent one tandem duplication in the ancestor of Neoteleostei and a second duplication (of sws2a) in the ancestor of Percomorpha (Cortesi et al. 2015). None of the studied species has two sws2a paralogs, suggesting that this gene was lost early in the evolution of Anabantiformes.

The green wavelength sensitive opsin *rh2* is the opsin that shows the highest variation in copy number along the teleost phylogeny, with copy numbers ranging from zero to eight (Musilova & Cortesi 2021). High numbers of *rh2* paralogs are mainly found in marine species inhabiting the open ocean (de Busserolles et al. 2020). Contrary to the high variation found in some lineages, we find little variation of *rh2* copy number within the Anabantiformes. While *C. argus* possess two *rh2* copies, the genomes of *A. testudineus* and of the five *Betta* species contain three *rh2* paralogs. We found two *rh2* copies in the genome of *H. temminckii*, both of which were on the border of two different scaffolds preventing
a reliable estimation of \( rh2 \) copy numbers in that species.

In contrast to the other cone opsin genes, the long-wavelength sensitive gene \( lws \) experienced an extensive amount of duplication events in the genomes of the labyrinth fishes. While the genomes of \( C. \ argus \), \( H. \ temminckii \) and \( A. \ testudineus \) contain two copies of this opsin gene, five paralogs were found in the genomes of all five \( Betta \) species. Even though the phylogeny of coding sequences of all genes suggests independent duplication events of \( lws \) in each lineage (\( Channa \), \( Helostoma \), \( Anabas \) and \( Betta \)), a synteny analysis of the region between \( lws1a \) and \( lws1b \) indicates that \( lws1a \) and \( lws1b \) of \( Betta \) are orthologs of the two \( lws \) copies in the other anabantiform species. This means that at least two tandem duplication events of \( lws \) must have happened in the lineage leading to \( Betta \), which is not in concordance with the results from Cortesi et al. (2021) suggesting one duplication in \( H. \ temminckii \) and \( A. \ testudineus \) and three duplications in the \( Betta \) lineage. Most teleost species only possess one or two copies of \( lws \) (Musilova et al. 2019), with the exception of a handful of lineages that experienced exceptional duplication of \( lws \) leading to up to five copies in some species (Cortesi et al. 2021). In some rare cases, the presence of multiple \( lws \) copies is the result of the ancient teleost specific genome duplications, but most species obtained their additional paralogs through relatively recent tandem duplications (Cortesi et al. 2021).

Using phylogenetic reconstructions to understand the evolutionary relationships of opsins proved to problematic, as gene conversion between opsin paralogs is common (Reyniers et al. 1995; Zhao et al. 1998; Hiwatashi et al. 2011) and can suppress differentiation between paralogs, but not between species. On the other hand, gene conversion can prevent pseudogenization of duplicated genes and even recover pseudogenized paralogs (Mighell et al. 2000; Cortesi et al. 2015), driving the increase of copy gene number and diversification between species, but decreasing within species differentiation at the same time.

Subfunctionalization through amino acid substitution

The high number of different opsin genes and paralogs of the same opsin class in some teleost species allows for nuanced fine-tuning of the visual apparatus to the different challenges brought by both the
biotic and abiotic environment. For this, different paralogs of the same opsin class must vary in their visual sensitivity, which is generally achieved through amino acid substitutions in the protein sequence of the opsin (Yokoyama 2008).

Our analysis of the amino acid sequence of LWS paralogs of the different Betta species revealed four regions corresponding to four of the seven transmembrane domains of LWS that seem to have diverged from the LWS sequence of other teleost species. LWS1a is the paralog most similar to other teleost lws genes, while LWS1b, LWS1c, LWS1d and LWS1e differ from them at the same positions. More importantly though, when comparing the LWS paralogs of Betta with each other, we see a very similar trend: LWS1a diverged from the other paralogs at the same position that diverged between Betta and other teleosts, while the remaining paralogs are relatively similar to each other. Again, the regions of differentiation correspond to four of transmembrane domains of the opsin, indicating functional differences between LWS1a and the other paralogs, as the transmembrane domains are the parts of the protein interacting with the chromophore and subsequently determine the spectral sensitivity of the opsin. This pattern of divergence suggests that LWS1a retained the spectral characteristics of LWS in the common ancestor of Anabantiformes, and LWS1b, LWS1c, LWS1d and LWS1e diverged to detect light at different wavelengths.

While in most cases it is not possible to make precise predictions about the changes of the wavelength of maximum absorption (\(\lambda_{\text{max}}\)) of an opsin in response to amino acid substitutions (for exceptions see Patel et al. (2018)), there are amino acid substitutions of five key sites of LWS that have been studied intensively (Yokoyama 2008; Yokoyama et al. 2008b). We found that two of these five positions are variable between the different LWS paralogs of Betta. The amino acid substitutions that we found indicate that \(\lambda_{\text{max}}\) of LWS1a is lower than the \(\lambda_{\text{max}}\) of the remaining paralogs and that the \(\lambda_{\text{max}}\) of LWS1b is lower than the \(\lambda_{\text{max}}\) of LWS1c, LWS1d and LWS1e. The presence of more amino acid substitutions affecting the \(\lambda_{\text{max}}\) of all five paralogs is likely, as four of the five key tuning sites are located within regions of amino acid divergence and spectral changes due to amino acid substitutions at key tuning sites can depend on the genetic background (Chinen et al. 2005). Even though, we cannot
precisely determine the spectral sensitive of the different LWS paralogs, our results of overall amino acid divergence and changes at key tuning sites suggest that there are functional differences between the LWS paralogs that allow the *Betta* species to detect light of different wavelengths.

Ontogenetic changes of opsin expression

Our analysis of opsin expression at eight different ontogenetic stages from the early larval stages to adulthood of *B. splendens* shows that each opsin gene found in the genome of *B. splendens* is expressed, at least at one ontogenetic stage. While the paralogs of all three cone opsins classes show considerable changes of expression, the relative expression between opsin classes remains relatively stable throughout the developmental stages, indicating that changes in spectral sensitivity between ontogenetic stages are rather achieved by switching between functionally diverse paralogs instead of changing the ratio of opsin classes, which are both common strategies in teleost fishes (Harer *et al.* 2017; Chang *et al.* 2021; Lupše *et al.* 2022). Paralogs with dominant expression in the early stage always are replaced by other paralogs during later developmental stages. Contrary to many other teleost species, that only express a small subset of cone opsin genes during adulthood (Spady *et al.* 2006; Harer *et al.* 2017), we found expression of 9 out of 10 existing cone opsin genes in adults of *B. splendens*.

The changes in expression levels of different paralogs of *lws* indicate a general shift of the spectral sensitivity from shorter wavelengths in the early developmental stages to longer wavelengths in later developmental stages, as the paralogs with the lower $\lambda_{\text{max}}$ are expressed early while paralogs with higher $\lambda_{\text{max}}$ are expressed during later developmental stages. This result is similar to observations in other teleost species where a general trend from short wavelength sensitivity in early developmental stages to long wavelength sensitivity in late developmental stages could be observed (Spady *et al.* 2006; Shand *et al.* 2008; Harer *et al.* 2017).

Temporal collinearity may facilitate subfunctionalization of *lws*

The expression of *lws* genes in *B. splendens* shows a very interesting pattern: paralogs located closer to the 5’end of the genomic cluster containing the five *lws* paralogs, start being expressed early during
development, while paralogs closer to the 3’end of the cluster start to be expressed later during
development, meaning that the relative position of the paralog in the genome corresponds to the
timing of expression during the life time of the animal. This phenomenon, called temporal collinearity,
is one of the most notable characteristics of the expression of hox genes during embryogenesis
(Duboule 1994). Even though it is not clear what causes the temporal shift of expression of hox genes,
it is known that changes of the chromatin structure make the downstream genes of the cluster
available for the transcription machinery during later stages of development (Noordermeer et al. 2014;
Deschamps & Duboule 2017). The temporal colinearity of lws genes might facilitate the
subfunctionalization of the paralogs of this opsin class. While sub- and neofunctionalization of opsin
genes rely on the functional diversification by changing the spectral sensitive of new paralogs (Spady
et al. 2006; Watson et al. 2011; Cortesi et al. 2021; Mitchell et al. 2021), the local (Torres-Dowdall et
al. 2017; Torres-Dowdall et al. 2021) but mostly the temporal (Spady et al. 2006; Carleton et al. 2008;
Harer et al. 2017; Chang et al. 2021) properties of gene expression must change for an advantageous
effect of the newly derived function. Given that under the assumption of colinearity, the temporal
expression of newly tandem duplicated genes is predetermined by their position in the gene cluster,
the only step for subfunctionalization is the change in spectral sensitivity. Even though our study does
not provide conclusive evidence for this hypothesis, temporal collinearity might be the cause for the
strong tendency of duplication and diversification of lws genes in Betta and other teleost lineages.
More support comes from the other opsin genes classes in Betta, where we found similar patterns of
collinearity. Even though our evidence is not conclusive, it provides a testable hypothesis for the
pattern of differentiation in opsin gene in teleost fishes.
Conclusion

The genomic analysis of different anabantoid species revealed that at least two duplication events occurred in the *Betta* lineage, resulting in the expansion of *lws* copy-number in this genus. Differences in amino acid sequences demonstrate functional differentiation between the different *lws* paralogs within *Betta*. The analysis of eye transcriptomes, revealed ontogenetic shifts of opsin expression. Temporal collinearity might be a factor driving *lws* duplication and diversification in *Betta*.

Acknowledgements

The authors thank Nadiya Orel, Femina Prabhukumar, Paolo Franchini and the staff of the animal research facility of the University of Konstanz for their valuable help.
Figure S 3.1 Phylogenetic reconstruction and recombination break points of \textit{lws}. Phylogenetic reconstruction of coding sequences of \textit{lws} upstream (base pairs 1-687) (A) and downstream (base pairs 688-1077) (B) of the recombination break point found using the Genetic Algorithm for Recombination Detection (GARD).
Figure S 3.2 Phylogenetic reconstruction of rh2 and sws. Phylogenetic reconstruction of coding sequences of rh2 (A) and sws2 (B). Green boxes indicate clades of rh2aα, rh2aβ and rh2aγ in Betta (A), blue boxes indicate clades of sws2a and sws2b in all acanthomorph species (B).
Figure S 3.3 Synteny dot plot of lws locus. Synteny dot plot comparing the similarity of the genomic region harboring lws1a and lws1b between C. argus and H. temminckii (A), C. argus and A. testudineus (B), H. temminckii and A. testudineus (C), B. splendens and A. testudineus (D), B. splendens and C. argus (E) and B. splendens and H. temminckii (F). Each black dot indicates two regions of 25bp with a similarity of at least 65%. Grey bars indicate the positions of exons of lws1a and lws1b.
Figure S 3.4 Phylogenetic reconstruction and recombination break points of \( rh2 \). Phylogenetic reconstruction of base pairs 1-158 (A), 159-701 (B) and 702-1056 (C) of the coding sequences of \( rh2 \) based on the recombination break points found using the Genetic Algorithm for Recombination Detection (GARD).
Figure S 3.5 Amino acid diversity of RH2 between Betta and other teleosts and within Betta. Sliding window analysis of Betta rh2 amino acid sequence divergence. Each green line represents the mean divergence between one rh2 paralog of all Betta species and rh2 sequences of a set of representative teleost species (A) or the mean divergence between one rh2 paralog of all Betta species and the remaining two paralogs in Betta (B). Grey areas indicate the position of transmembrane domains of rh2.
Figure S 3.6 Amino acid diversity of SWS2 between Betta and other teleosts and within Betta. Sliding window analysis of Betta sws2 amino acid sequence divergence. Each green blue represents the mean divergence between one sws2 paralog of all Betta species and sws2 sequences of a set of representative teleost species (A). The black line represents the mean divergence between the two sws2 paralogs of all Betta species (B). Grey areas indicate the position of transmembrane domains of sws2.
Table S 3.1 Estimated maximal absorption of lws genes in B. splendens. Estimated shift of maximal absorption between rh2 genes in B. splendens based on amino acid substitutions at sites of one known key tuning sites of rh2 (Yokoyama et al. 2008b).

<table>
<thead>
<tr>
<th>Tuning site</th>
<th>Bo. taurus</th>
<th>B. splendens</th>
<th>Betta splendens</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>rh2aa</td>
</tr>
<tr>
<td></td>
<td>122</td>
<td>130</td>
<td>E</td>
</tr>
<tr>
<td>Estimated shift of $\lambda_{max}$ (nm)</td>
<td>$+10$</td>
<td>$0$</td>
<td>$+10$</td>
</tr>
</tbody>
</table>
General discussion

While the classical idea of adaptive evolution was based on the idea of mutations in protein-coding regions of the genome that would result in amino acid changes which subsequently lead to the formation of new phenotypes, toady it is well established that variation in gene regulation plays an important role in the fast diversification of phenotypes. Changes in gene regulation can result in higher or lower amounts of expression (Kratochwil et al. 2018) but also in temporal and spatial changes of gene expression (e.g. in the visual system of different teleost species (Harer et al. 2017; Torres-Dowdall et al. 2021) and Chapter 3). One clear difference in comparison to changes in the coding region is that changes in gene regulation can have fine-tuned, specific effects, while changes in the coding sequence of the gene will most probably change the function in all contexts and might hence have pleiotropic effects on multiple phenotypes (Cooper et al. 2011). In contrast, changes in gene regulation can be restricted to certain time points during ontogeny or certain locations within the organism. On the other hand, changes in gene regulation are not able to change the actual function of the gene product, which may be a necessary response to changes in the environment (Tufts et al. 2015).

It has been shown that changes in gene regulation can promote fast phenotypic diversification in different adaptive radiations. In cichlids divergent gene expression can cause divergence along multiple phenotypic axis including color patterns (Kratochwil et al. 2018; Kratochwil et al. 2022), jaw morphology (Parsons et al. 2014; Ahi et al. 2019b), aggression (Baran & Streelman 2020), fin shape (Ahi et al. 2017; Ahi et al. 2019a), breeding behavior (York et al. 2018) and the visual system (Hofmann & Carleton 2009; Harer et al. 2018). Furthermore, changes in gene regulation drove phenotypic diversification e.g. in beak shapes of Darwin’s finches (Abzhanov et al. 2004; Abzhanov et al. 2006), color patterns in lepidopteran butterflies and moths (Reed et al. 2008; van’t Hof et al. 2016; McMillan et al. 2020) and wing phenotypes in Drosophila (Koshikawa et al. 2015), clearly demonstrating the important role of gene regulation changes in the formation of phenotypic diversity in a wide range of taxa.
The multitude of different methodological approaches used in the presented thesis, demonstrated that phenotypic diversity in closely related teleost species can have a variety of different sources. The mechanisms underlying the phenotypic diversity in the studied species include differences in regulation of gene expression, copy number variation and mutations in the protein-coding region of genes that subsequently lead to amino acid substitutions of the final protein.

Gene regulation and the formation of phenotypic diversity
In chapter 1 of this thesis, we could demonstrate that phenotypic diversity can be driven by or associated with changes in gene expression. In this study, the formation of horizontal stripes in Lake Malawi cichlids has been associated with multiple genomic loci, but most notably with a large effect QTL encompassing the gene *agrp2*. Previous studies demonstrated a strong association between low expression levels of *agrp2* in the skin of adult individuals and the presence of horizontal stripes in East African cichlids (Kratochwil et al. 2018) and other teleost species (Liang et al. 2021). The results of our study suggest that the presence of horizontal stripes in the studied species *Pseudotropheus cyaneorhabdos* is driven by the regulation of *agrp2*, but further indicate a more complex regulation of the formation of the phenotype as we found multiple large effects QTL affecting the formation of horizontal stripes. The evolution of novel regulatory mechanisms to drive the formation of stripe patterns does not come as a surprise, as it has been demonstrated before that mechanisms to change expression levels of *agrp2* in Lake Malawi and Lake Victoria cichlids evolved independently (Urban et al. 2020), indicating that the regulation of stripe pattern formation evolves dynamically. Knowledge about the temporal and spatial dynamics of *agrp2* expression and its effects on horizontal stripe formation is scarce. While striped and non-striped regions of the skin of African cichlids do not show differences in *agrp2* expression, the striped portion of the skin of the Neotropical cichlid species *Trichromis salvini* and the zebrafish *Danio rerio* show significantly lower expression levels of *agrp2* when compared to the non-striped regions, demonstrating the possible role of spatial regulation of
gene expression on the diversification of phenotypes.

The results of chapter 1 could further demonstrate that the formation of vertical bars is driven by multiple genomic loci. In contrast to horizontal stripes, we could not identify a single large effect QTL and we could not find obvious candidate genes that might influence the formation of color patterns within the identified genomic intervals. Previous studies investigating the development of vertical bar patterns in the Lake Victoria species *Haplochromis latifasciatus* found differential expression of coloration related genes between barred and non-barred regions (Liang et al. 2020) suggesting a role of differential gene expression in the formation of this color pattern. For a more detailed understanding of the genetic basis of vertical bars, it would be helpful to further hybridize individuals of the F$_2$ generation or hybridize individuals of the F$_1$ generation with non-barred parental species. The many small effect loci influencing the formation of vertical bars would be isolated and easier to detect in a study using association mapping (Tanksley & Nelson 1995; Bernacchi et al. 1998). In chapter 2, we investigated the genetic basis of aggressive behavior of the fighting fish species *B. splendens* and *B. simplex*. Our transcriptomic analysis demonstrated that different levels of aggression in the *Betta* species were associated with individual, fine-tuned regulatory responses targeting the short-term but also long-term regulation of social behavior. The changes in both species underlie differential expression of various genes, none of them outstanding on their own, showing the complexity of genetic regulation of social behavior. Complex transcriptional changes in the brain as a response to a social interaction have been found in a number of teleost species. Similar to our results, Bukhari et al. (2017) demonstrated that individuals of three-spined stickleback (*Gasterosteus aculeatus*) show strong transcriptional changes in the diencephalon and telencephalon in response to a territorial challenge. While many of the differentially expressed genes had functions related to the regulation of behavior it was not possible to point out single candidate genes with outstanding effects on the behavioral response. Instead, it was argued that changes in the expression of transcription factors and chromatin accessibility are important drivers of the transcriptional response.

In contrast to our results that show a high number of up- and down-regulated genes, other
studies investigating the transcriptomic response to social interactions in *B. splendens* found that a majority of differentially expressed genes are upregulated (Vu *et al.* 2020). A major difference between the studies was the use of whole brain tissue in Vu *et al.* (2020) as opposed to the use of specific brain regions in chapter 2 of this thesis. The use of whole brain tissue may not be able to detect region specific expression changes as these might be covered by overall higher expression in the other brain regions, emphasizing the importance of the spatial regulation of gene expression. Gene expression changes that alter the behavior may be limited to small brain regions or even single neurons (van Heukelum *et al.* 2019; Guthman & Falkner 2022) and might be hard to detect when using whole brain tissue for the transcriptomic analysis. A possible way to refine the differential expression analysis in the context of social behavior would be the use of single-cell RNA sequencing. After identifying brain regions or single neurons that show high activity during an aggressive encounter (Bahl & Engert 2020) the analysis of the transcriptomes of these cells might give detailed insights into expression changes (Siddique *et al.* 2021) leading to changes in behavior. This approach might be able to provide a more conclusive answer to the questions how changing gene expression can drive changes and diversification of social behavior taking into account the spatial, cell-specific regulation of gene expression.

Gene duplication and gene regulation
Another way to circumvent the problem of pleiotropic effects can be the rise of new genetic material by gene duplication (Ohno 1970). While in most cases duplicated genes lose their function due to the release of selective pressure (Lynch & Conery 2000), it is also possible that the new copy of a gene is retained and evolves functionally (Rastogi & Liberles 2005). The whole genome duplication that happened in the common ancestor of all teleost species (Meyer & Van de Peer 2005), resulted in the expansion of many gene families and is one of the factors contributing to the successful diversification of teleost fishes (Peatman & Liu 2007; Lu *et al.* 2012; Rennison *et al.* 2012).
Interestingly, diversification after gene duplication cannot only happen by means of mutations in the coding regions of the new paralogs, but again through changes of regulation of expression of the new gene copies. In Chapter 3, the important role of gene duplication in the creation of phenotypic diversification through expression regulation is emphasized. The duplication of the red-sensitive genes \( lws \) in the genus \( Betta \) drove the diversification of different paralogs as demonstrated by high levels of amino acid divergence between paralogs and the estimated absorption spectrum of the different paralogs. As demonstrated by the expression of the cone-opsin genes in \( B. splendens \) during ontogeny, paralogs that diverged functionally can further separate by changing their timing of expression. Accordingly, cone-opsin genes of all classes switch between expression of the different paralogs during the development of the fish. Temporal changes of cone-opsin expression are well documented in teleosts (Lupše et al. 2022) and have been connected to changing feeding habits throughout the lifetime of organisms (Jordan et al. 2004; Hofmann et al. 2009; O'Quin et al. 2010) or coloration-based mate choice (Seehusen et al. 2008; Sandkam et al. 2018). Diversification of spatial expression has been found in Neotropical cichlids (Torres-Dowdall et al. 2021) leading to different spectral sensitivity of different parts of the retina. Our results show, that changes in gene regulation can promote the evolution of phenotypic diversity in combination with other evolutionary processes.

Conclusion
This thesis demonstrates the important role of changes in gene regulation on the evolution of rapid phenotypic diversification. Using different methodological approaches in different model organisms, we could demonstrate how changes in gene regulation can drive diversity of body coloration, social behavior and the visual system of two different teleost radiations in combination with other evolutionary processes such as hybridization and gene duplication. Further, we demonstrate the need to use a wide range of methodological approaches and study systems to achieve a comprehensive understanding of the genetics of phenotypic diversification. The diverse genetic routes that can lead
to the evolution of new phenotypes can only be detected using an equally diverse set of molecular methods and a diverse set of model organisms.
Author contributions

Chapter 1: Of bars and stripes: a Malawi cichlid hybrid cross provides insights into genetic modularity and evolution of modifier loci underlying color pattern diversification
CFK, JG and AM conceptualized the project. JG and SU collected the data. JG analyzed the data. CFK and JG wrote the manuscript. SU and AM revised the manuscript.

Chapter 2: Brain transcriptomics provide insights into short- and long-term regulation of aggressive behavior in the Siamese fighting fish, *Betta splendens*
JG and AM conceptualized the project. JG collected and analyzed the data. JG wrote the manuscript.

Chapter 3: Seeing red: evolution of cone opsin genes in the Siamese fighting fish, *Betta splendens*
JTD and JG conceptualized the project. JG collected the data. JG and JTD analyzed the data. JG and JTD wrote the manuscript.
Acknowledgements

Even though only my name can be found on the title page of this thesis, the completion of it would not have been possible without the help and effort of many other people. I want to take the opportunity to thank all of you.

Obwohl nur mein Name auf dem Titelblatt dieser Dissertation zu lesen ist, wäre ihre Fertigstellung ohne die Hilfe und den Einsatz vieler anderer Menschen nicht möglich gewesen. Ich möchte diese Gelegenheit nutzen, euch zu danken.

At first, I want to thank the members of my thesis committee Prof. Axel Meyer, Prof. Catherine Peichel and Prof. Daniel Dietrich. Axel, thank you for the opportunity to conduct my doctoral studies in your lab and for the opportunity to meet and work with a collective of great scientists. Prof. Peichel, Prof. Dietrich, thank you for sacrificing some of your precious time to read my thesis and grade my work. I deeply appreciate it.


Julián, I want to thank you for the guidance you provided throughout the past years. From my very start in the lab as Master student, you always spared some of your precious time and energy to give feedback on or brainstorm for projects you were not even involved in. I am happy that we actually found a project to work on as a team during the past two years and I am grateful for everything I learned from your during that time.


In einem großen Labor gibt es ständig neues zu entdecken und zu lernen. Lieber Joost, lieber Paolo, lieber Andi, lieber Ralf, vielen Dank, dass ihr euer Wissen mit mir geteil habt.


Mein Dank gilt außerdem allen Mitgliedern der IMPRS, insbesondere den Koordinatorinnen Mäggi Hieber Ruiz und Francisca Rosa Mende. Vielen Dank für die moralische aber auch die finanzielle Unterstützung.

Ein besonderer Dank gilt der Hector Fellow Academy für die Finanzierung meiner Arbeit der letzten 4 Jahre.

Lieber Pascal, liebe Vanessa, vielen Dank, dass ihr mich in den letzten 10 Jahren daran erinnert habt, dass es auch außerhalb der Universität noch ein Leben gibt. Ihr habt für mich Konstanz zu einem Stück Heimat gemacht.

123

Publications


References


56.


57.


58.


59.


60.


61.


62.


63.


64.


65.


66.


67.


79.

80.


81.


82.


83.


84.


85.


86.


87.


88.


89.


90.


91.

92.


93.


94.


95.


96.


97.


98.


99.


100.


101.


102.


103.


104.


155. 137


170.


171.


172.


173.


174.


175.


176.


177.


178.


179.


180.


181.


182.


208.


209.


210.


211.


212.


213.


214.


215.


216.


217.


218.


219.


220.

221.


222.


223.


224.


225.


226.


227.


228.


229.


230.


231.


232.


233.


146


