Molecular basis of a novel pigment trait in cichlid fishes

Inauguraldissertation

zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel
von

Maria Emília Pombo dos Santos
von Portugal

Basel, 2014

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät
Auf Antrag von

Prof. Dr. Walter Salzburger, Dr. Patrícia Beldade

Basel, 18 September 2012

Prof. Dr. Jörg Schibler
The dean of faculty

Table of Contents

Abstract 1

Chapter 1: Introduction 2

Chapter 2: Comparative transcriptomics of Eastern African Cichlid fishes shows signs of positive selection and a large contribution of untranslated regions to genetic diversity 16

Published research manuscript

Chapter 3: The evolution of cichlid egg-spots is linked with a *cis*-regulatory change 30

In press research manuscript

Chapter 4: Transcriptomics of a novel and variable pigment trait in cichlid fishes – Identification of candidate genes for egg-spot morphogenesis 55

4.1 Abstract 57

4.2 Background 58

4.3 Results 61

4.4 Discussion 73

4.5 Conclusions 75

4.6 Methods 76

4.7 Supplementary information 84

4.8 Acknowledgements 85

4.9 References 86

Chapter 5: How cichlids diversify 90

Published perspective manuscript

Chapter 6: The ecological and genetic basis of convergent thick-lipped phenotypes in cichlid fishes 95

Published research manuscript

Chapter 7: Discussion and future perspectives 111

Acknowledgements 116

Curriculum Vitae 117

Abstract

The genetics underlying the evolution of novel morphological structures is a fascinating topic that has attracted the attention of many evolutionary biologists. Among the East African cichlid fauna, the haplochromines represent the most species-rich group. One of their characteristics is the occurrence of egg-spots on the anal fins of males, which mimic real eggs and play a crucial role in the breeding cycle of these maternal mouthbrooding fish. These yellow to orange egg-spots serve as intra-specific sexual advertisement to attract females and to maximize breeding success. They are a novel trait that emerged only once in the evolution of the haplochromine lineage. The main goal of this doctoral thesis was to deepen our understanding of the genetics and developmental basis of the emergence and diversification of egg-spots, an evolutionary novelty in East African cichlid fishes. Further understanding of the molecular basis novelty requires the identification of the genes and mutations that underlie these major phenotypic changes. Here we report the identification of two genes that are involved in the development of the egg-spot trait – *fhl2a* and *fhl2b* – and one possible *cis*-regulatory mutation in *fhl2b* that might have played a role in the emergence of the egg-spot trait. We further described many more candidate genes via an RNAseq survey of *Astatotilapia burtoni* (haplochromine species) egg-spot and anal fin transcriptome. We generated hypotheses about their possible function using Gene Ontology definitions and inter-species gene expression, establishing a database that will serve as an important resource and useful resource for future research on the emergence and diversification the egg-spot trait.

Chapter 1

Introduction

Introduction

How novel morphological structures evolve is one of the most fascinating topics in evolutionary biology[1–3]. Some of the famous examples of novelty include the emergence of flowers in angiosperms [4, 5], the evolution of insect wings [6, 7], the presence of horns in beetles [8, 9], butterfly eyespots [10], the shells of turtles [11], the vertebrate neural crest [12], and the evolution of eyes [13]. Novel traits are fascinating by themselves because they are examples of the extremely diverse and astonishing range of phenotypes that evolution is able to create. Furthermore, novel traits represent discontinuities in the phenotypic range of traits and therefore attract the attention of many biologists. How does a novel trait emerge? What are the genetic and developmental mechanisms underlying the origin of novel traits? Recently, these topics have become the focus of research in evolutionary biology. In this thesis I set out to further understand the mechanisms underlying the emergence of novel traits, focusing on the characteristic egg-spots that emerged in East African cichlid fishes.

What is a morphological novelty?

What exactly is a "novel trait" is a matter of controversy and many definitions have been proposed. Ernst Mayr [14] defined a novelty as ''any newly acquired structure or property that permits the performance of a new function, which in turn will open a new adaptive zone''. This concept directly links novel traits to adaptation and, also, to instances of adaptive radiation. There are some novelties, however, that are not connected with radiations and unfairly fall out of the scope of novelty according to this definition. Müller and Wagner defined novelty as ''…a structure that is neither homologous to any structure in the ancestral species or homonomous to any other structure in the same organism'' [15]. The authors try to set a homology threshold, i.e. novelty would begin where homology would end. This is a very stringent definition and the boundaries of homology are rather ill defined. In a recent essay, Pigliucci suggested that "Evolutionary novelties are new traits or behaviours, or novel combinations of previously existing traits or behaviors, arising during the evolution of a lineage, and that perform a new function within the ecology of that lineage" [16]. This definition has the advantage of not implying any mechanism for the origin of novelties (contrary to the homology definition) and not implying that the new function is correlated with an adaptive radiation. How novelties emerge and what are the developmental and genetic mechanisms that underlie its origins remain unknown. As a consequence the mechanism of origin should not be used to define what is a novel trait. The last definition of novelty has the problem of taking us back to the grey area of quantitative and qualitative differences: How different does one trait have to be in order to be considered novel? Clearly, the existing definitions of novelty cannot reach an agreement and are still controversial. Importantly, there will only be a consensus when we understand how novelties originate, and what the grey area between variation of a pre-existing trait and novel trait is. Knowledge of how a trait comes into existence and how it is maintained will greatly facilitate our understanding of what evolutionary novelty is.

Mechanisms underlying novelty

The origin of a novel trait requires the emergence of a new developmental module that will give this trait its unique identity. Recently it has been suggested that this new developmental program might result from the recruitment of pre-existing genetic networks, where “old” genes perform new “tricks” [17–19]. The origin and development of horned beetles, for example, is connected with the recruitment of limb patterning genes [20] and, similarly, limb development in vertebrates has been connected to Hox gene co-option [21]. There is consensus that changes in gene expression underlie the recruitment of these networks, but what causes the change is still debated [22, 23]. Changes in gene expression can either result from changes in the *cis*-regulatory region, or modifications in the protein sequence of transcription factors [24, 25]. For example, it was shown that a novel wing colour pattern in *Drosophila guttifera* involves the co-option of new expression sites of the Wingless morphogen, and that this co-option is due to a *cis*-regulatory change [26]. On the other hand there are cases where the difference in gene expression patterns that contribute to the origin of novel traits relies on changes in the protein sequence of transcription factors that will ultimately activate the expression of downstream targets. For example, the emergence of pregnancy in placental mammals is thought to have been accompanied by a change in the interaction dynamics between Homeobox A11 (*HoxA11*) and Forkhead box O1A (*Foxo1a*), which play a major role in regulating gene expression [27]. Both *cis*-regulation and protein evolution seem to play a role in the emergence of novel traits, although which mechanism is the main mode of change in gene expression (if there is one) is still unknown. The evolutionary history of co-option of pre-existing genes and networks also raises some questions. Is the gene network pre-wired before co-option or are its component genes co-opted into this new developmental network one at a time [28]? Recent work done in butterfly species reveals that this process is not so simple, suggesting that both scenarios are possible and stressing the importance of broader phylogenetic studies in order to understand the evolutionary history of gene and gene network recruitment [29].

Co-option of pre-existing mechanisms seems to be the driving force in morphological evolution. Recently, though, there has been accumulating evidence that new lineage specific genes can play a role in the development of novel traits [30, 31]. For example, studies in Hydra (*Hydra magnipapillata*) showed that new lineage specific genes played a role in the evolution of novel traits such as the cnidarian nematocyte [32]. There are many more open questions than clear answers concerning the emergence of novelty and the genetics of this process is largely unknown. Clearly, more case studies are needed in order to understand how novelties emerge and diversify.

Addressing the origin of novelty

Addressing the genetic basis of the origin of a novel trait is a difficult task. Detecting genetic variation responsible for the phenotypic variation in novel trait is feasible, but the genes that underlie the variation might not be the ones that were responsible for the emergence of the novelty. The alternative is to take a comparative approach and compare the development of the trait between lineages that possess the novel trait with ancestral lineages that do not. The novel trait should represent the biggest morphological change in the smallest timescale possible, so that the species we are comparing should be as closely related as possible. By understanding the development of the novel traits and its genetics, we might find the genes and pathways underlying the trait and then be able to disentangle novelty-coincident and novelty-causative changes through functional assays.

Egg-spots: a cichlid fish evolutionary novelty

Egg-spots are colorful circular markings on the anal fins of many hundreds of cichlid fishes. They are a novel trait characteristic of the most species rich cichlid lineage – the haplochromines, which are a major part of the East African cichlid diversity (figure 1). East African cichlid fishes, including the hundreds of endemic species found in lake Malawi and lake Victoria, are the result of the most spectacular adaptive radiations known in vertebrates and provide an ideal system to study the molecular basis of evolutionary novelties in the context of adaptation and explosive speciation.

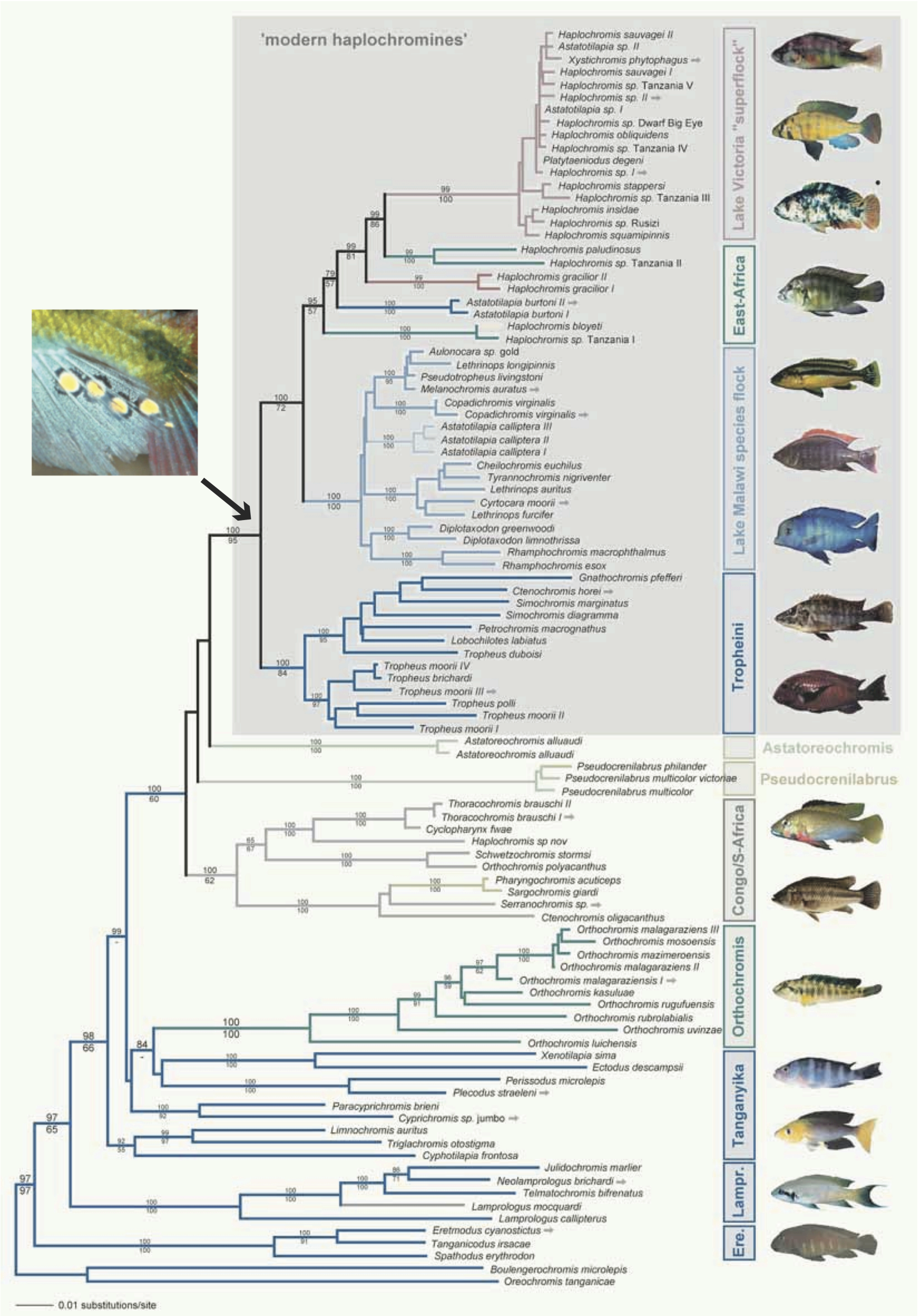


Figure 1 – Representative phylogeny of East African Cichlids (modified from [39]). Emergence of the egg-spots is coincident with the emergence and radiation of the modern haplochromine clade and is signaled with an arrow.

They are the most species-rich vertebrate group (~2000 species) and, despite their close relatedness, they show extreme phenotypic diversity [33–37]. A representative phylogeny of East African cichlids is depicted in figure 1. Haplochromines are the most diverse and species-rich lineage of cichlids, harbouring approximately 1500 species. This group represents up to 80% of East African cichlid diversity and can therefore be considered the most successful cichlid lineage [37–39]. Haplochromines sometimes show extreme sexual dimorphism where males are large and extremely colourful, with females being smaller and dull in coloration [40, 41]. Egg-spots are a putative key innovation in the haplochromines and its evolutionary origin coincides with the origin of the modern haplochromine lineage (figure 1). It has been suggested that egg-spots have promoted the haplochromines' astonishing diversification and speciation [38, 42, 43].

Egg-spots morphology and function

Egg-spots are present on the anal fin of haplochromine males and consist of a central circular area of xanthophores surrounded by an outer transparent ring. They are highly variable in colour, number and arrangement, both within and between species (figure 2). Egg-spots can be yellow, orange or red, and can vary in number from one to dozens depending on the individual and species. They have been suggested to mimic real eggs and hence are sometimes referred to as "egg-dummies" [44].

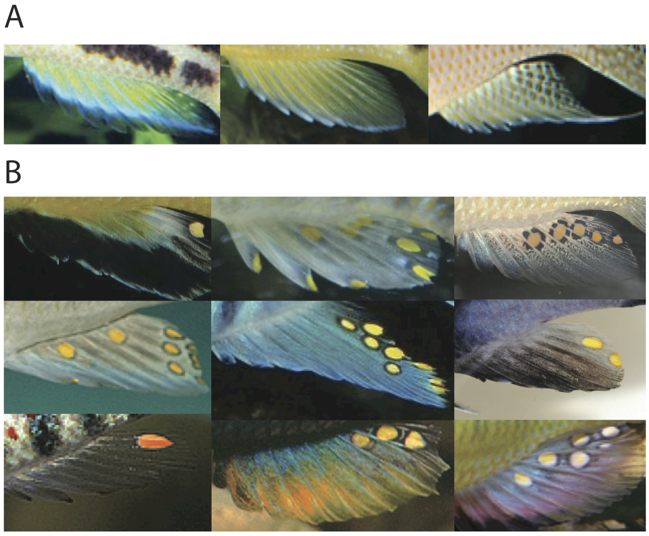
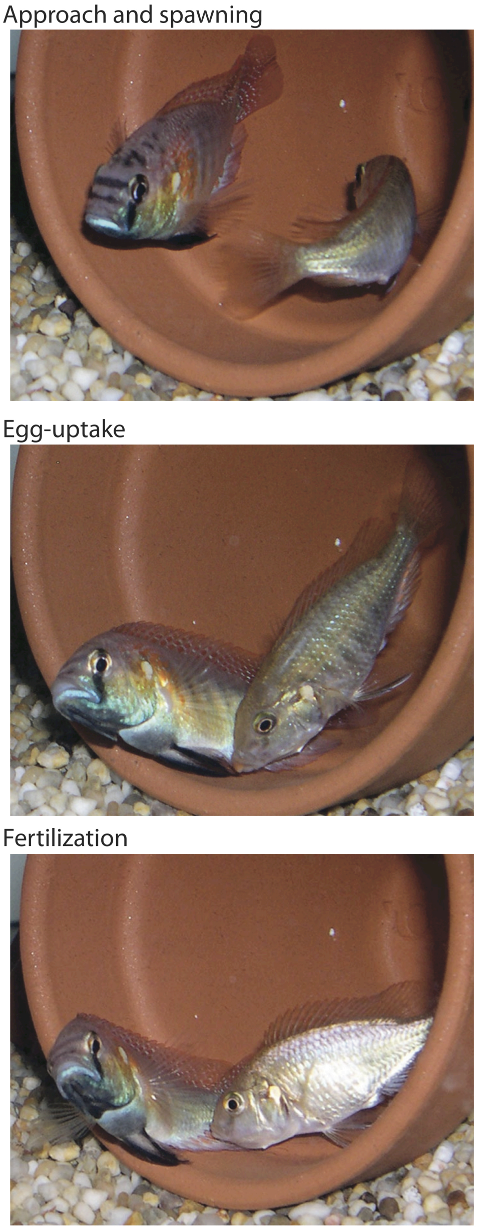


Figure 2 - Anal fins from several East African cichlids **A)** non-haplochromine anal fins with no egg-spots **B)** haplochromine anal fins with egg-spots. (Pictures from unknown source).

Haplochomine females are mouthbrooders, meaning that the females brood and carry their young in their buccal cavity. Egg-spots appear to play a role in the courtship and spawning behaviour of these maternal mouthbrooding species. During this behaviour, the male laterally displays his colourful fins whilst the female approaches the male's territory. The female then lays a batch of eggs in the male's territory and, before fertilization takes place, picks them up in her mouth. The female then interacts with the egg-spots on the anal fin of the male in a behaviour that seems like she is trying to bite them, at this moment the male discharges his sperm and fertilization takes place in the female's mouth (figure 3) [40, 41]. The exact function of the egg-spot signal across species seems to be as variable as their morphology. In *Astatotilapia elegans*, egg-spots play a role in female choice, where females prefer males with more egg-spots [45]. A similar scenario was observed in *Pseudotropheus aurora*, where females spawned more frequently with males showing more egg-spots, with the number of egg-spots correlating with the number of clutches fertilized. Therefore, males with more egg-spots had higher fitness [46]. These two studies led to the conclusion that egg-spots serve as intra-specific sexual advertisement and to maximize breeding success. In *Pseudotropheus lombardoi*, egg-spots appear to play a different role. The males of this species show only one spot, and females prefer males with one spot over males where another egg-spot had been artificially added [47]. This indicates that in this species this trait might be linked to species recognition. Recent studies in *Astatotilapia burtoni* show that males with more egg-spots are the dominant males and that the egg-spots serve as an important signal in intra-sexual male-male aggression and competition



Approach and spawning

Egg-uptake

Fertilization

Figure 3 – Courtship behaviour of the haplochromine *Astatotilapia burtoni*. The female approaches the male territory where she will lay a batch of eggs that she will pick up in her mouth before total fertilization takes place. The female then interacts with the egg-spots in the anal fin of the male and it is in this moment that the sperm is discharged. Fertilization takes place in the female's mouth.

for territories [48, 49]. All these studies demonstrate that egg-spots are a sexually selected trait, either via female choice, or via male-male competition, and in addition they might function as a signal for species recognition.

Developmental origin of egg-spots

Egg-spots are mainly made up of a central circular area of xanthophores surrounded by an outer transparent ring [50]. Vertebrate pigment cells derive from a migratory cell lineage - the neural crest [51, 52]. The neural crest is a pluripotent cell line that delaminates from the embryonic neural tube and adjacent ectoderm. These cells then migrate through different routes giving rise to different vertebrate traits, including neuron cells, pigment cells, craniofacial bones, Schwann cells, and smooth muscle cells [51]. In fish, the neural crest produces six different pigment cells – melanophores (black or brown), xanthophores (yellow or red), erythrophores (red), leucophores (white), cyanophores (blue) and iridophores (reflective/iridescent). Adult pigmentation is a result of the differential migration, survival, proliferation, and interaction of these pigment cells [53]. Egg-spots start to develop during the transition between the juvenile and adult stage, together with other sexually dimorphic traits [54]. Three main processes seem to be involved in the development of the egg-spots and anal fin pigmentation: neural crest differentiation and migration, cell migration-adhesion cues that will pattern the egg-spots in the anal fin, and finally pigment production. The interaction and differences between these processes will result in the astonishing diversity we see in the haplochromine egg-spots.

Genetic basis of egg-spots

The study of the genetics underlying the egg-spot phenotype is in its infancy. So far only one paper has been published on the topic showing that the xanthophore marker colony stimulating factor 1 receptor A (*csf1ra*) is involved in the formation of egg-spots of several haplochromine species [50]. The authors analyzed the coding region of this gene in several cichlid lineages and found that it underwent adaptive sequence evolution in the direct ancestral lineage leading to the modern haplochromines. This fact suggests that *csf1ra* might play an important role in the evolution of egg-spots. This gene (csf1ra) is involved on the onset of xanthophore pigment production and is therefore involved in one of the downstream processes of egg-spot formation. In order to understand the origin and evolution of this novel phenotype, we need to address more upstream genes.

Aim: Understanding the genetics of emergence and diversification of egg-spots

Egg-spots are a haplochromine cichlid novel trait thought to have facilitated the diversification of this lineage via speciation through sexual selection. It serves as an ideal system to understand the genetics of the emergence and further diversification of novel traits. The egg-spot represents a dramatic morphological change in pigmentation and, since cichlid species are very closely related, egg-spots provide the ideal phylogenetic framework for comparative studies that will help understand the evolution of novel traits. The aim of this doctoral thesis is to generate a better understanding of the genetics of egg-spot emergence and diversification.

Thesis outline

The main goal of my thesis was to find and characterize candidate genes for egg-spot development in order to advance our knowledge on the genetic basis of this novel trait. Egg-spots are not a well-established novelty model system, and so far only one gene has shown to be correlated with its development. Therefore, as a first step, I was heavily involved in a project to generate two transcriptome datasets, one from a haplochomine species (*Astatotilapia burtoni*) and one from an ectodine species (*Ophtalmotilapia ventralis*). Well-characterized transcriptomes are important sequence resources that can greatly aid the identification of genes underlying phenotypic variation by further transcriptomic experiments (eg. RNAseq). Chapter two is the resulting paper from this work describing this process and the results of the analysis of these two species' transcriptomes.

In chapter three, I conducted an RNAseq experiment comparing female and male fins in order to generate candidate genes involved in the egg-spot trait. I followed up on the two most male biased expressed genes, *fhl2a* and *fhl2b,* and confirmed that these are indeed involved in haplochromine egg-spot development. I then found evidence that a *cis*-regulatory mutation upstream of *fhl2b* might have contributed to the emergence of this trait.

In chapter four I further characterized the egg-spot transcriptome by comparing fins within individuals, with the aim of finding groups of candidate genes. I confirmed many of these genes as egg-spot genes and generated hypotheses about their possible function using Gene Ontology [55] definitions and inter-species gene expression comparisons. With these chapters I have established the egg-spot as model trait for the study of novelties and generated many candidates that will be useful for future studies.

The last two chapters are tangents to the main study that I was involved in during my time as a PhD student. Chapter five is a perspective on the cichlid model system in light of the recent release of genome sequences for five species. Cichlids are a very popular system in speciation and adaptive radiation research, although not as popular in developmental biology. This perspective was written to demonstrate the potential of the model as a whole and to elucidate the different areas of evolutionary biology research that we can now tackle with cichlids. Chapter six is a published paper, on which I am a co-author, investigating the genetics of convergent cichlid thick-lip phenotypes. Finally, in chapter seven, I discuss the results obtained throughout the doctoral work, along with brief suggestions for future directions of study.

References

1. Moczek AP, Sultan S, Foster S, Ledón-Rettig C, Dworkin I, Nijhout HF, Abouheif E, Pfennig DW: **The role of developmental plasticity in evolutionary innovation.** *Proceedings. Biological sciences / The Royal Society* 2011, **278**:2705-13.

2. Wagner GP, Lynch VJ: **Evolutionary novelties.** *Current Biology* 2010, **20**:R48-52.

3. Müller GB, Newman SA: **The Innovation Triad: An EvoDevo Agenda**. 2005, **304B**:487-503.

4. Stebbins GL: **Adaptive Radiation of Reproductive Characteristics in Angiosperms, I: Pollination Mechanisms**. *Annual Review of Ecology and Systematics* 1970, **1**:307-326.

5. Albert VA, Oppenheimer DG: **Pleiotropy, redundancy and the evolution of flowers**. *Trends in Plant Science* 2002, **7**:297-301.

6. Averof M, Cohen S: **Evolutionary origin of insect wings from ancestral gills**. *Nature* 1997, **385**:627-630.

7. Jockusch EL, Nagy LM: **Insect evolution: how did insect wings originate?** *Current Biology* 1997, **7**:R358-361.

8. Moczek AP, Rose D, Sewell W, Kesselring BR: **Conservation, innovation, and the evolution of horned beetle diversity.** *Development genes and evolution* 2006, **216**:655-665.

9. Emlen DJ, Corley Lavine L, Ewen-Campen B: **On the origin and evolutionary diversification of beetle horns.** *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**:8661-8668.

10. Beldade P, Brakefield PM: **The genetics and evo-devo of butterfly wing patterns.** *Nature Reviews Genetics* 2002, **3**:442-452.

11. Rieppel O: **Turtles as hopeful monsters.** *BioEssays* 2001, **23**:987-991.

12. Shimeld SM, Holland PW: **Vertebrate innovations.** *Proceedings of the National Academy of Sciences of the United States of America* 2000, **97**:4449-4452.

13. Fernald RD: **Casting a genetic light on the evolution of eyes.** *Science* 2006, **313**:1914-1918.

14. Mayr E: *Animal Species and Evolution*. Cambridge, MA: Harvard University Press; 1963.

15. Müller GB, Wagner GP: **Novelty in Evolution: Restructuring the Concept**. *Annual Review of Ecology and Systematics* 1991, **22**:229-256.

16. Pigliucci M: **What, if Anything, Is an Evolutionary Novelty?** *Philosophy of Science* 2008, **75**:887-898.

17. True JR, Carroll SB: **Gene co-option in physiological and morphological evolution.** *Annual Review of Cell and Developmental Biology* 2002, **18**:53-80.

18. Carroll SB, Grenier JK, Weatherbee SD: *From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design*. 1st edition. Oxford, UK: Blackwell Science; 2001.

19. Carroll SB: **Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution.** *Cell* 2008, **134**:25-36.

20. Moczek AP, Rose DJ: **Differential recruitment of limb patterning genes during development and diversification of beetle horns.** *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106**:8992-8997.

21. Hérault Y, Beckers J, Duboule D, Gérard M: **Hox Gene Expression in Limbs: Colinearity by Opposite Regulatory Controls**. *Developmental Biology* 1999, **208**:157-165.

22. Alonso CR, Wilkins AS: **The molecular elements that underlie developmental evolution**. *Nature Reviews Genetics* 2005, **6**:709-715.

23. Stern DL, Orgogozo V: **The loci of evolution: how predictable is genetic evolution?** *Evolution* 2008, **62**:2155-2177.

24. Prud'homme B, Gompel N, Carroll SB: **Emerging principles of regulatory evolution.** *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**:8605-8612.

25. Lynch VJ, Wagner GP: **Resurrecting the role of transcription factor change in developmental evolution.** *Evolution* 2008, **62**:2131-2154.

26. Werner T, Koshikawa S, Williams TM, Carroll SB: **Generation of a novel wing colour pattern by the Wingless morphogen.** *Nature* 2010, **464**:1143-1148.

27. Brayer K, Lynch VJ, Wagner GP: **Evolution of a derived protein–protein interaction between HoxA11 and Foxo1a in mammals caused by changes in intramolecular regulation**. *Proceedings of the National Academy of Sciences of the United States of America* 2011, **108**:E414–E420.

28. Monteiro A, Podlaha O: **Wings, horns, and butterfly eyespots: how do complex traits evolve?** *PLoS Biology* 2009, **7**:e37.

29. Shirai LT, Saenko SV, Keller RA, Jerónimo MA, Brakefield PM, Descimon H, Wahlberg N, Beldade P: **Evolutionary history of the recruitment of conserved developmental genes in association to the formation and diversification of a novel trait.** *BMC Evolutionary Biology* 2012, **12**:21.

30. Martin A, Reed RD: **Wingless and aristaless2 define a developmental ground plan for moth and butterfly wing pattern evolution.** *Molecular Biology and Evolution* 2010, **27**:2864-2878.

31. Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TCG: **More than just orphans: are taxonomically-restricted genes important in evolution?** *Trends in Genetics* 2009, **25**:404-13.

32. Milde S, Hemmrich G, Anton-Erxleben F, Khalturin K, Wittlieb J, Bosch TCG: **Characterization of taxonomically restricted genes in a phylum-restricted cell type.** *Genome Biology* 2009, **10**:R8.

33. Stiassny MLJ, Meyer A: **Cichlids of the Rift Lakes**. *Scientific American* 1999, **280**:64-69.

34. Kocher TD: **Adaptive evolution and explosive speciation: the cichlid fish model.** *Nature Reviews Genetics* 2004, **5**:288-298.

35. Seehausen O: **African cichlid fish: a model system in adaptive radiation research.** *Proceedings of the Royal Society B* 2006, **273**:1987-1998.

36. Salzburger W: **The interaction of sexually and naturally selected traits in the adaptive radiations of cichlid fishes.** *Molecular Ecology* 2009, **18**:169-185.

37. Turner GF, Seehausen O, Knight ME, Allender CJ, Robinson RL: **How many species of cichlid fishes are there in African lakes?** *Molecular Ecology* 2001, **10**:793-806.

38. Salzburger W, Mack T, Verheyen E, Meyer A: **Out of Tanganyika: genesis, explosive speciation, key-innovations and phylogeography of the haplochromine cichlid fishes.** *BMC Evolutionary Biology* 2005, **5**:17.

39. Verheyen E, Salzburger W, Snoeks J, Meyer A: **Origin of the superflock of cichlid fishes from Lake Victoria, East Africa.** *Science* 2003, **300**:325-329.

40. Fryer G, Iles T: *The Cichlid Fishes of the Great Lakes of Africa: Their Biology and Evolution*. Edinburgh, UK: Oliver & Boyd; 1972.

41. Barlow GW: *The Cichlid Fishes: Nature's Grand Experiment in Evolution*. 1st edition. Cambridge, MA: Perseus Publishing; 2000.

42. Wagner CE, Harmon LJ, Seehausen O: **Ecological opportunity and sexual selection together predict adaptive radiation**. *Nature* 2012, **487**:1-5.

43. Goldschmidt T, de Visser J: **On the possible role of egg mimics in speciation**. *Acta Biotheoretica* 1990, **38**:125-134.

44. Wickler W: **"Egg-dummies" as natural releasers in mouth-breeding cichlids**. *Nature* 1962, **194**:1092–1093.

45. Hert E: **The function of egg-spots in an African mouth-brooding cichlid fish**. *Animal Behaviour* 1989, **37**:726–732.

46. Hert E: **Female choice based on egg-spots in Pseudotropheus aurora Burgess 1976, a rock-dwelling cichlid of Lake Malawi, Africa.** *Journal of Fish Biology* 1991, **38**:951–953.

47. Couldridge V: **Experimental manipulation of male eggspots demonstrates female preference for one large spot in Pseudotropheus lombardoi**. *Journal of Fish Biology* 2002, **60**:726-730.

48. Lehtonen TK, Meyer A: **Heritability and adaptive significance of the number of egg-dummies in the cichlid fish Astatotilapia burtoni.** *Proceedings of the Royal Society B* 2011, **278**:2318-2324.

49. Theis A, Salzburger W, Egger B: **The function of anal fin egg-spots in the cichlid fish Astatotilapia burtoni.** *PloS One* 2012, **7**:e29878.

50. Salzburger W, Braasch I, Meyer A: **Adaptive sequence evolution in a color gene involved in the formation of the characteristic egg-dummies of male haplochromine cichlid fishes.** *BMC Biology* 2007, **5**:51.

51. LaBonne C, Bronner-Fraser M: **Molecular mechanisms of neural crest formation**. *Annual Review of Cell and Developmental Biology* 1999, **15**:81-112.

52. Meulemans D, Bronner-Fraser M: **Gene-regulatory interactions in neural crest evolution and development.** *Developmental Cell* 2004, **7**:291-299.

53. Fujii R: **The regulation of motile activity in fish chromatophores.** *Pigment Cell Research* 2000, **13**:300-319.

54. Heule C, Salzburger W: **The ontogenetic development of egg-spots in the haplochromine cichlid fish Astatotilapia burtoni.** *Journal of Fish Biology* 2011, **78**:1588-1593.

55. Ashburner M, Ball C, Blake J: **Gene Ontology: tool for the unification of biology**. *Nature Genetics* 2000, **25**:25-29.

Chapter 2

Comparative transcriptomics of Eastern African Cichlid fishes shows signs of positive selection and a large contribution of untranslated regions to genetic diversity

Laura Baldo, M. Emília Santos, Walter Salzburger

Published in Genome Biology and Evolution

Personal contribution:
In this study I contributed to the study design, sample handling, sequencing and manuscript preparation

Comparative Transcriptomics of Eastern African Cichlid Fishes Shows Signs of Positive Selection and a Large Contribution of Untranslated Regions to Genetic Diversity

Laura Baldo*, M.Emília Santos, and Walter Salzburger*

Zoological Institute, University of Basel, Basel, Switzerland

*Corresponding author: E-mail: laura.baldo@unibas.ch;walter.salzburger@unibas.ch.

Data deposition: Assembled contigs were deposited in GenBank under accession numbers JL478463-JL524178b for Astatotilapia burtoni and JL554673 - JL597291 for Ophthalmotilapia ventralis. All sequence alignments used in this study are available from the corresponding author upon request.

Accepted: 9 May 2011

Abstract

The hundreds of endemic species of cichlid fishes in the East African Great Lakes Tanganyika, Malawi, and Victoria are a prime model system in evolutionary biology. With five genomes currently being sequenced, eastern African cichlids also represent a forthcoming genomic model for evolutionary studies of genotype-to-phenotype processes in adaptive radiations. Here we report the functional annotation and comparative analyses of transcriptome data sets for two eastern African cichlid species, *Astatotilapia burtoni* and Ophthalmotilapia *ventralis*, representatives of the modern haplochromines and ectodines, respectively. Nearly 647,000 expressed sequence tags were assembled in more than 46,000 contigs for each species using the 454 sequencing technology, largely expanding the current sequence data set publicly available for these cichlids. Total predicted coverage of their proteome diversity is approximately 50% for both species. Comparative qualitative and quantitative analyses show very similar transcriptome data for the two species in terms of both functional annotation and relative abundance of gene ontology terms expressed. Average genetic distance between species is 1.75% when all transcript types are considered including nonannotated sequences, 1.33% for annotated sequences only including untranslated regions, and decreases to nearly half, 0.95%, for coding sequences only, suggesting a large contribution of noncoding regions to their genetic diversity. Comparative analyses across the two species, tilapia and the outgroup medaka based on an overlapping data set of 1,216 genes (~526 kb) demonstrate cichlid-specific signature of disruptive selection and provide a set of candidate genes that are putatively under positive selection. Overall, these data sets offer the genetic platform for future comparative analyses in light of the upcoming genomes for this taxonomic group.

Key words: *Astatotilapia burtoni*, Ophthalmotilapia *ventralis*, positive selection, EST, 454 sequencing, UTR.

Introduction

Cichlid fishes from eastern African Great rift lakes and surrounding rivers represent a major model for rapid speciation in evolutionary biology (Kocher 2004; Seehausen 2006; Salzburger 2009). More than 1,500 endemic species have arisen in a few millions of years only, showing the most spectacular adaptive radiations known in vertebrates (Seehausen 2006). Explosive radiations in the cichlid species flocks of lakes Victoria, Malawi, and Tanganyika are mostly documented by paleo-geographical (i.e., the ages of the lakes) and molecular data. Lake Victoria, for example, is only between 200,000 and 500,000 years old and fell dry about 15,000 years ago (Johnson et al. 1996). Still, it harbors an endemic flock of several hundred species that are likely to have diversified in a maximum of about 100,000 years only (Verheyen et al. 2003). Accordingly, preliminary molecular data from partial genomes, nuclear and mitochondrial markers of East African cichlids have inferred a highly similar genetic background among species (Sturmbauer and Meyer 1993; Aparicio et al. 2002; Loh et al. 2008). This is in strong contrast with their tremendous diversity of morphotypes and ecological adaptations (Salzburger 2009) suggesting that, in cichlids, rapid phenotypic diversification is largely uncoupled from an equivalent molecular diversity in coding regions. Hence, cichlids represent an ideal system to dissect the genetic bases of several universal phenotypic traits (such as coloration, body morphology, color vision, etc.) and—more generally speaking—to explore the molecular evolutionary processes underlying diversification and ecological speciation.

An increasing number of studies in animals points to the diversity of transcriptomes and especially of the expression profiles (thus including regulation of gene expression) as the bridging link that translates highly similar genomes at protein-coding genes into the astonishing diversity of phenomes (i.e., set of phenotypes) (see, e.g., Cooper et al. 2003; Wray et al. 2003; Shapiro et al. 2004). In particular, regulatory changes involving a limited genetic diversity can affect the expression of alternatively spliced isoforms and may modulate timing, localization, and abundance of gene expression. These processes can be adaptive and, therefore, responsible for organismal diversification (reviewed by Fay and Wittkopp 2008).

To date, comparative transcriptome analyses of African cichlids have been limited in terms of species number and number of expressed sequence tags (ESTs) analyzed (Salzburger et al. 2008; Kobayashi et al. 2009; Lee et al. 2010). These studies, overall, revealed a high uniformity of the protein-coding sequences among closely related, yet phenotypically diverse species.

Here, we report more than a million new EST sequences, perform transcriptome analyses, and investigate the overall expression profiles of two African cichlid species, *Astatotilapia burtoni* (AB) and Ophthalmotilapia *ventralis* (OV). AB and OV are representatives of two main evolutionary cichlid lineages (tribes) from East Africa, the modern haplochromines and the more basal group of the ectodines, respectively (see, e.g., Salzburger et al. 2002, 2005). The two lineages are thought to have diverged several millions of years ago (Salzburger et al. 2005; Koblmuller et al. 2008). So far, comparative genetic studies between these two lineages were largely limited to a phylogenetic context (see, e.g., Salzburger et al. 2002, 2005; but see Salzburger et al. 2007), whereas genomic comparisons are lacking. The two species differ in body morphology, ecology, and behavior. AB is a mouth-brooding species found in rivers and estuaries around Lake Tanganyika and is characterized by the presence of "true" circular egg-spots on the anal fins of males. OV is also a mouth-brooding species endemic to lake Tanganyika but exhibits long ventral fins showing egg-dummies in form of yellow vessels at their tips. Functional egg-dummies are, hence, a feature that evolved several times during cichlid evolution in East Africa (Salzburger et al. 2007; Salzburger 2009).

For each of these two species, more than 647,000 ESTs were generated through 454 sequencing (Roche) and assembled in more than 46,000 contigs. These represent the first 454 data sets and the largest collection of EST available to date for African cichlids. This study also provides the first transcriptome data for a member of the ectodine lineage (OV). Functional annotation and comparative analyses were performed to explore major qualitative and quantitative differences of the two transcriptomes. Furthermore, comparative analyses were expanded to include additional species via the identification of more than a 1,200 orthologoues contigs across AB, OV, the Nile tilapia (*Oreochromis niloticus*) and medaka (*Oryzia latipes*) as outgroup. This allowed screening for differential substitution rates along lineages and for individual genes. Overall, our study provides an important molecular resource for comparative studies within cichlids and among fishes in general and will facilitate the assembling and annotation of the upcoming cichlid genomes (http://www.broadinstitute.org/models/tilapia).

Materials and Methods

Samples

Specimens from an inbreed laboratory strain of AB were kept at the University of Basel (Switzerland) under standard laboratory conditions. OV individuals were captured live in Mpulungu (Zambia), shipped to Basel, and kept at the same laboratory conditions for a week. For RNA isolation, individuals were euthanized with MS 222 using approved procedures (permit nr. 2317 issued by the cantonal veterinary office).

cDNA Library Construction and 454 Sequencing

From AB, we extracted total RNA from ten embryos, ten fish larvae, two juveniles, and two adults (one male and one female). From OV, we used four adults (three males and one female). For each species, specimens were pooled together, roughly chopped, and incubated for 2 h in 8 ml of trizol (Invitrogen). Samples were then ground to complete homogenization using a mortar and a pestel. RNA extraction was performed according to the manufacturer's protocol. DNase treatment was carried out with the DNA-free Kit (Applied Biosystems). The quantity and quality of RNA were assessed by spectrophotometry and gel electrophoresis. One microgram of RNA of each sample was sent for commercial normalized library construction by Vertis Biotechnology AG (http://www.vertis-biotech.com/). From total RNA, first strand cDNA was synthetized using a reverse transcriptase, an N6 random primer and a small aliquot of an oligo(dT)-primer for enrichment of 3′ ends. 454 adapters A and B were ligated to the 5′ and 3′ ends of the cDNA. cDNAs were then amplified by polymerase chain reaction (PCR) (15 cycles) using a proofreading enzyme. Libraries were normalized by hydroxyl-apatite chromatography, and the single-stranded cDNA was amplified by PCR (nine cycles). cDNA was then selected with gel fractioning for fragments of sizes 500 to 700 bp.

Normalized cDNA libraries for the two species were sequenced with a Roche Genome Sequencer FLX system

(Roche 454) in one Titanium FLX run (two lanes, one for each species) by Microsynth (http://www.microsynth.ch). Base calling was performed with Phred (Ewing et al. 1998). Reads were assembled with the GS De Novo Assembler version 2.0.0.22 using the default settings, a minimum overlap of 40 nucleotides and identity threshold of 90%.

ESTs Functional Annotation

Gene ontology (GO) annotation was conducted using Blast2GO version 2.4.4 (Conesa et al. 2005). Briefly, BlastX searches were performed against the nonredundant database (nr) using the QBlast for multiple queries, setting the *e* value to 1.0 × 10^{-6}, the high scoring segment (HSP) length cut off greater than 33 and the number of hits to 5. GO annotation was done using the following settings: a pre-*E*-value-Hit-Filter of 1.0 × 10^{-6}, a GO weight of 5, and the annotation cut off of 55. Contigs with no significant hits to the nr data set were BlastN searched against the nucleotide database (nt) for possible identification, setting the expected cut off value to 1.0 × 10^{-15}.

Clustering of Orthologous Sequences

For the purpose of obtaining a data set suitable for comparative analyses, we generated three data sets, which included orthologous ESTs across AB and OV (data set #1), AB, OV, and *O. niloticus* (hereafter referred to as tilapia) (data set #2), and AB, OV, tilapia, and *O. latipes* (hereafter referred to as medaka) (data set #3). Data set #3 represented a subset of data set #2.

For the data set #1, identification of orthologous ESTs between the two species was performed using a bidirectional best hit (BBH) method (Overbeek et al. 1999). Reciprocal batch BlastN searches were carried out setting the expected value cut off to 1.0 × 10^{-50} to minimize significant matches to paralogous sequences. Outputs were analyzed using in-house R scripts. Hits with a bit score > 1,000 were retrieved for further analyses. Pairwise assemblies were performed using CodonCode Aligner version 3.7.1 (Codon Code Corporation) and aligned with MAFFT version 6.821b (Katoh et al. 2002) using a local pairwise method based on the Smith–Waterman algorithm.

For data set #2, a total of 117,222 tilapia ESTs were downloaded from GenBank in September 2010 (Lee et al. 2010). Among the total BBHs, we selected only annotated BBHs that had a length overlap > 400 bp and a bit score > 400. Contigs from both AB and OV belonging to this subset were batch BlastN searched against the tilapia data set, setting the expected value cut off value to 1.0 × 10^{-50}. Corresponding best hits for the two species to the tilapia data set that had a length overlap > 150 bp were retrieved, assembled in CodonCode Aligner and aligned in MAFFT. Alignments were trimmed for full-length overlap.

Finally, for data set #3, all contigs belonging to the data set #2 (2,660) from AB were batch BlastX searched against complete protein data sets from *Danio rerio* and medaka (retrieved from the ENSEMBL database v59) using a cutoff of 1.0 × 10^{-50}. Significant hits with concordant frames between *D. rerio* and medaka were chosen, and the corresponding cDNA sequences from medaka were retrieved from ENSEMBL. Clusters of orthologues cDNA sequences across medaka and the three cichlid species were generated and aligned using MAFFT. *Danio rerio* sequences were not included due to the high nucleotide divergence of this species with respect to the other species (Steinke et al. 2006). To obtain only open reading frames (ORFs), untranslated regions (UTRs) were trimmed from the alignments according to the corresponding medaka proteins. All frame-shifting indels introduced in Medaka sequences during the aligning process were trimmed to preserve medaka-reading frames. Alignments below 150 bp in length were discarded. Finally, all alignments were eye checked and refined manually.

The final data set #3 comprised 1,216 alignments of fully overlapping sequences starting with the correct reading frames. The pipeline was performed with in-house R and perl scripts.

Phylogeny, Genetic Distances, and Rates of Evolution

Maximum likelihood (ML) heuristic searches were performed on the concatenated alignment of 1,216 four-species clusters (526,113 bp) from data set #3 using RaxML version 7.0.4 (Stamatakis et al. 2005). We performed a rapid bootstrap analysis and search for the best ML tree employing the GTRGAMMA model. Indels were identified using the program SeqState (Muller 2005). All single and double indels present in cichlid sequences in the final alignments (36 and 5, respectively) were considered as sequencing errors and replaced with Ns. Two deletions longer than 100 bp identified in OV were attributed to a putative exon skipping (alternative spliced variants) and not to a genomic deletion and also replaced with Ns. Indels were then coded using the simple indel coding strategy (Simmons and Ochoterena 2000), implemented in SeqState, and mapped on the ML tree performing a maximum parsimony analysis in PAUP* v. 4.0b10 (Swofford 2000).

Uncorrected distance matrices were estimated for individual alignments using PAUP*. Pairwise synonymous and nonsynonymous substitution rates per site (Ks and Ka, d*S* and d*N*) were estimated under two methods; the Nei and Gojobori method (Nei and Gojobori 1986) implemented in the DNAStatistics package of Bioperl (http://www.bioperl.org/wiki/Main_Page) (Ks and Ka) and the Goldman and Yang method (Goldman and Yang 1994) using the program Codeml implemented in PAML version 4.4b (Yang 2007) (d*S* and d*N*).

Different rates of d*N*/d*S* for branches in the phylogenetic tree were investigated using the branch models from Codeml. d*N*/d*S* values were averaged across sites (NSsites = 0). Three models of molecular evolution were compared: 1) the one-ratio model (model = 0), allowing the same d*N*/d*S* value for all branches; 2) the two-ratio model, constraining the branches within the cichlid clade to one d*N*/d*S* ratio that was different from all the others (model = 2); and 3) the free-ratio model (model = 1), allowing one d*N*/d*S* ratio per each branch. Sites with ambiguous data were removed (cleandata = 1). The three models were compared (0 vs. 2 and 2 vs. 1) using a likelihood ratio test (LRT) with two and four degrees of freedom, respectively.

Positive selection acting on genes that showed average Ka/Ks values higher than one between species was further tested by estimating d*N*/d*S* for branches in individual gene phylogeny under the free-ratio model in Codeml.

Results

ESTs Sequence Annotation and Comparative Transcriptomics of AB and OV

The two EST libraries constructed for AB and OV yielded an equal number of reads (~647,000), which were assembled in a similar number of contigs (>46,000, see table 1). The mean contig size was 585 bp for AB and 566 bp for OV, with 39% of the contigs having at least 500 bp.

Based on BlastX searches against the nr database, 19,121 AB (38.8% of the total) and 16,585 OV (35.8%) contigs had a significant hit above the cut off *e* value of 10^{-6} (table 2). These contigs corresponded to a total of 12,491 distinct accession numbers (AccNos) for AB and 11,269 AccNos for OV. Because the contigs are usually much shorter than the corresponding cDNA sequences, it is common that several contigs matched to the same gene, in spite of lacking adequate overlap to be assembled. For both species, the top-hit species for orthologue match was *Tetraodon nigroviridis* (approximately 35% of the contigs), followed by *D. rerio* (approximately 25%).

Of the contigs with significant BlastX hits, a total of 11,956 for AB and 10,250 for OV were annotated in 4,852 GO terms (24% of the total contigs) and 5,152 GO terms (22%), respectively. The GO terms were assigned to three biological categories that were equally represented in the two species (table 2). Relative and absolute abundance of the most represented GO terms per biological category were also comparable between AB and OV (fig. 1). The two species shared nine of ten terms in all three categories. The most represented terms for the molecular function category were associated to protein and nucleotide binding and transcription factor activity, whereas the predominant terms for the biological process category were involved in common enzymatic processes such as “auxin biosynthetic process,” “oxidation reduction,” and “signal transduction”. Finally, overrepresented GO terms for the cellular component category were mainly localized in the nucleus and membrane.

A large part of the contigs had no significant hit to the nr data set (above 60% for both species). These contigs were BlastN searched against the nt database for further identification. Only 9% of these contigs for both species (2,863 and 2,620 contigs for AB and OV, respectively) returned a significant hit to the nt database (1×10^{-15}), with 609 unique AccNos shared between the two species (see supplementary table S1, Supplementary Material online). Of these AccNos, several (up to 100) mapped to noncoding regions, such as microsatellite sequences, pseudogenes, and transposons. We also retrieved genes predicted to play an important role in cichlid evolution, such as *Bmp4*, *c-ski*, *pax* genes, prolactin, *Sox* transcription factors, the vitellogenin receptor, among others. In terms of frequency of contigs per single hit, half of the total number of contigs mapped to the same two classes of genes in both species and with similar relative proportions (table 3): immune genes (MHC class, KLR, natural killer-like receptors), and patterning genes (Hox and ParaHox genes). This suggests that both a relatively high expression of these genes in the two species, as well as poor amino acid conservation outside the cichlid lineage that could explain why these contigs did not return any BlastX hit against the nr database. To some extent, this outcome might also be biased by the overrepresentation of these loci in GenBank.

Table 1
Summary of the ESTs Generated by 454 Sequencing in This Study

	AB	OV
Summary run		
Total number of reads	647,219	647,816
Average read length	349.27	344.36
Total number of bases	226,048,424	223,072,738
Summary assembly		
Total number of contigs	49,311	46,298
Total number of large contigs (≥500 bases)	19,408	17,207
Average contig size	585.84	566.33
N50 contig size[a]	1,016	1,003
Largest contig size	8,335	7,430

[a] Half of all bases reside in contigs of this size or longer.

Comparative Transcriptomics within Cichlids

Using the BBH method, we identified 20,828 contigs that had best reciprocal hits between AB and OV. Of these, a total of 4,516 contigs that had a BlastN score bit ≥ 1,000 were selected to explore sequence diversity between the species (data set #1). These clusters of putatively orthologous sequences comprised a representation of all transcript types, such as annotated and nonannotated sequences, as well as coding and noncoding regions (including UTRs). The average alignment length was 1,463 bp with a mean pairwise nucleotide distance, excluding indels, of 0.0175 ± 0.0101, and a median of 0.0158 (table 4).

Table 2
Summary of the ESTs Annotation Using Blast2GO

	AB	OV
Number of ESTs returning BlastX hits	19,121 (12,491 AccNos)	16,582 (11,269 AccNos)
Number of ESTs with GO annotation	11,956 (5,152 terms)	10,250 (4,852 terms)
Biological process	8,438 (2,974 terms)	7,293 (2,732 terms)
Cellular component	7,330 (616 terms)	6,307 (623 terms)
Molecular function	10,110 (1,562 terms)	8,683 (1,497 terms)
Annotated protein-coding genes	8,684	7,671

Considering only annotated sequences, we generated 2,660 clusters of orthologous contigs among AB, OV and tilapia (data set #2) that could reliably be aligned. Average pairwise genetic distance was virtually the same between tilapia and both AB and OV (~0.030) and more than twice as large as between OV and AB (0.0138) (table 4). Genetic distance between AB and OV was higher than the one calculated in the previous data set, likely because this second

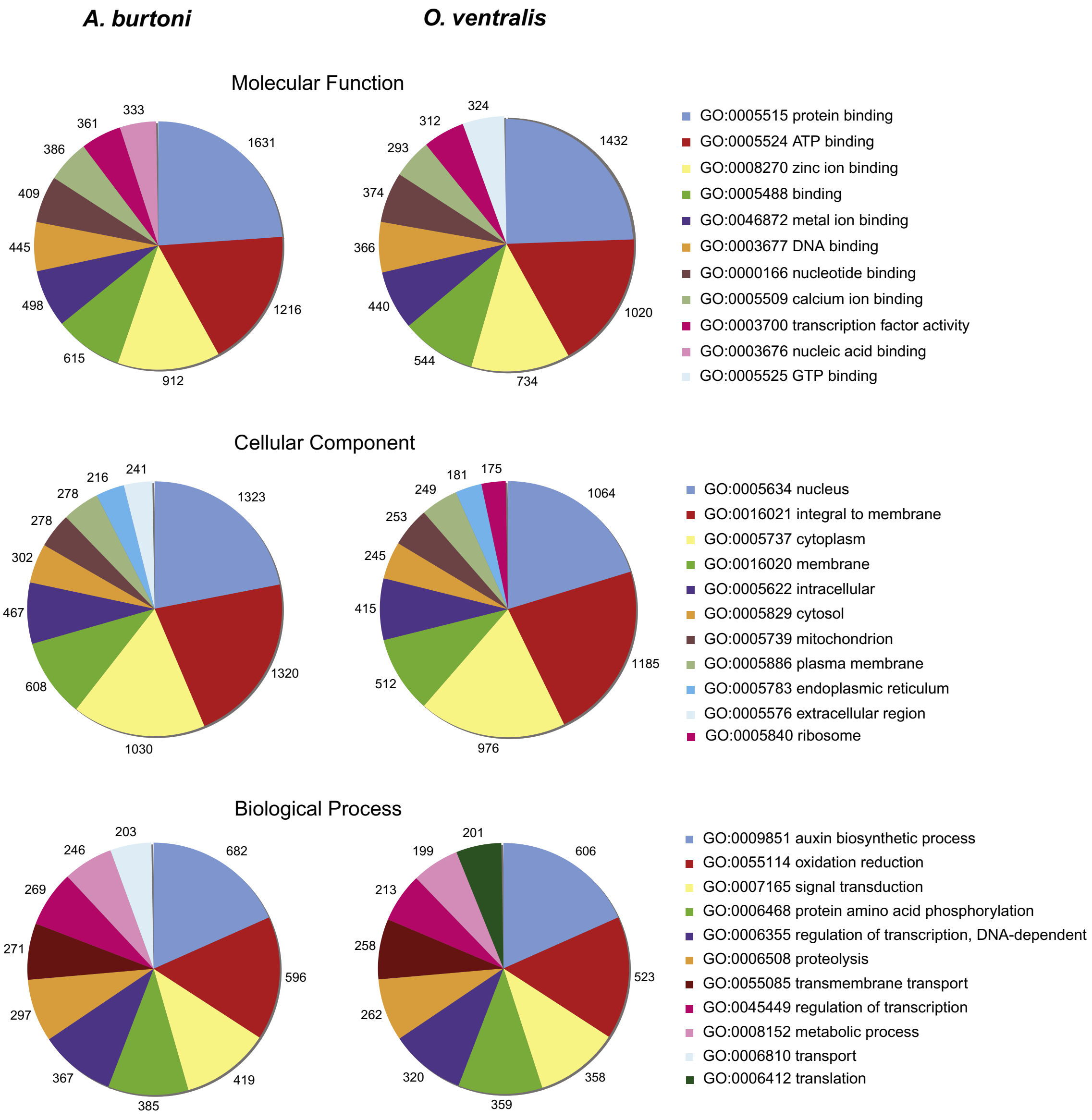


Fig. 1.—Ten most represented GO terms per biological category and absolute number of ESTs assigned to each term. Overall representation of GO terms is nearly equal between AB and OV.

Table 3
Most Common Hits in the nt Database (cut off *e* value 1 × 10^{-15}) for Contigs That Had No Hits in the nr Database

Hit Description	Species	AccNo	Number of Contigs	
			AB	OV
MHC class IA antigen UBA1, UBA2, UAA1 genes, UAA3 and UAA2 pseudogenes, UAA4, UAA5, and UAA6 pseudogene fragments	*Oreochromis niloticus*	AB270897.1	260	226
Platelet-derived growth factor receptor beta b (pdgfrbb) and colony-stimulating factor 1 receptor b (csf1rb) genes	*Astatotilapia burtoni*	DQ386647.1	181	153
Hoxba gene cluster	*A. burtoni*	EF594310.1	149	136
KLR1 gene; KLR2 pseudogene, KLR3 and KLR4 genes; KLR5 gene, KLR6 and KLR7 pseudogenes	*O. niloticus*	AY495714.1	115	115
Hoxdb gene cluster	*A. burtoni*	EF594316.1	84	59
Platelet-derived growth factor receptor beta a (pdgfrba) and colony-stimulating factor 1 receptor a (csf1ra) genes	*A. burtoni*	DQ386648.1	60	43
Gsh2 (gsh2), Pdgfra (pdgfra), and Kita (kita) genesKdrb (kdrb) gene; and Clock (clock) gene	*A. burtoni*	EF526075.2	57	64
Hoxbb gene cluster	*A. burtoni*	EF594314.1	56	74
Hoxab gene cluster, complete sequence	*A. burtoni*	EF594311.1	55	52
KLR8 pseudogene; KLR9 gene, C-type lectin (CLECT2)-like protein pseudogene, and C-type lectin (CLECT2)-like protein gene; KLR10 pseudogene; C-type lectin natural killer cell receptor-like protein gene; and transposon TX1-like ORF2 pseudogene	*O. niloticus*	AY495715.1	45	47
Hoxda gene cluster	*A. burtoni*	EF594315.1	31	32
Hoxca gene cluster	*A. burtoni*	EF594312.1	22	30
Hoxaa gene cluster	*A. burtoni*	EF594313.1	20	13
Total number of contigs			1,135	1,044

data set only included annotated sequences, thus excluding all novel, less conserved, and untranslated mRNA sequences (but yet including UTR regions).

We finally generated a third data set (#3) including orthologous sequences across the three cichlid species and the outgroup medaka. UTRs were trimmed using medaka proteins as reference. We obtained 1,409 clusters of fully overlapping orthologous sequences across AB, OV, tilapia, and medaka that included only ORFs. Inspection of the alignments revealed 191 clusters in which premature stop codons were present in one or more cichlid species but not in medaka. These stop codons could represent sequencing errors or real substitutions resulting in pseudogenization or truncation of proteins (with potential novel functions). At this stage, we could not tease apart the three scenarios and we therefore decided to exclude these clusters from the data set. The final data set #3 comprised 1,216 four-species alignments of ORFs, with a total length of 526,113 bp. Average length for individual alignments was 433 bp, varying between 153 and 741 bp. We used this data set for phylogenetic reconstructions and to investigate genetic diversity and levels of selection for each species pairwise comparison and along phylogenetic lineages.

Table 4
Average Pairwise Genetic Distance (Pi, Uncorrected) with Standard Deviation and Median Values Estimated from 4,516 BBHs between AB and OV (Data set #1) and from 2,660 Three-Species Alignments (AB, OV, and Tilapia; Data set #2)

		Pi	Median	Mean Length (Range), bp
Data set #1[a]				
AB	OV	0.0175 ± 0.0101	0.0158	1,463 (516–6,837)
Data set #2				
AB	OV	0.0138 ± 0.0096	0.0117	541 (150–2,588)
Tilapia	AB	0.0302 ± 0.0203	0.0261	
Tilapia	OV	0.0314 ± 0.0212	0.0268	

[a] Data set #1 includes both annotated and nonannotated ESTs, whereas data set #2 includes only annotated ESTs with UTRs.

The ML phylogeny based on the concatenated data set is shown in figure 2. The tree is in accordance with previously reported phylogenetic relationships among the four species (Salzburger et al. 2005; Steinke et al. 2006): AB and OV grouped together and formed a well-supported monophyletic group with tilapia (bootstrap values = 100 for both nodes). The three cichlids showed similar genetic distance from the outgroup medaka.

In accordance with the phylogenetic reconstruction, the shortest absolute genetic distance was found between AB and OV (0.0095), followed by tilapia versus these two species (0.0222 and 0.0230), with the longest distance occurring between medaka and the remaining three species (0.1605 and 0.1609) (table 5). Within cichlids, contribution of indels to the genetic diversity was low, with a total of 268 indel sites detected out of 524,047 nucleotides. These corresponded to a total of 38 distinct indel events equally

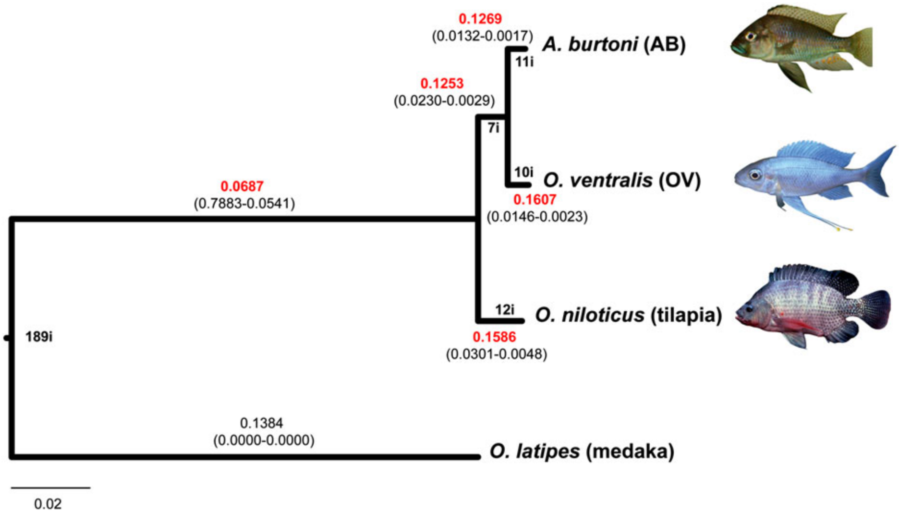


Fig. 2.—ML phylogeny based on four-species concatenated alignment of 1,216 genes (526,113 bp). The tree is rooted using medaka as outgroup. All nodes had a 100 bootstrap value support. For each branch, individual d*N* and d*S* values (in brackets, respectively) and the corresponding d*N*/d*S* ratios (in red) were calculated under the free-ratio model (codeml). Indel events per branch (specified by number followed by "i") were mapped by maximum parsimony.

distributed along the three cichlid lineages (AB, OV, and tilapia) (mapped in fig. 2). Six deletion events (3- to 6-bp deletions) were specific of the AB/OV clade and occurred in the following genes: the "low choriolytic enzyme precursor," involved in the breakdown of the egg envelope, the "src kinase-associated phosphoprotein 2," involved in the src signaling pathway, the "deoxyribonuclease tatdn3," the "v-type atpase b subunit," the "dna topoisomerase 2-beta," and the "probable rna-binding protein eif1ad." Further investigations are needed to clarify whether these amino acid deletions confer important biological changes to these proteins and are therefore involved in some cichlid-specific traits.

UTRs Contribution to Cichlid Genetic Divergence

To check for the specific contribution of UTRs to the genetic diversity between cichlid species, pairwise genetic distances were calculated on the same gene data set as data set #3 (thus excluding nonannotated sequences) before and after trimming for fully coding sequences (table 6). The average length of the 1,216 alignments among the three cichlid species including partial or full UTRs was 536 bp, ranging between 156 and 1,746 bp, for a total of 652,849 bp. Inclusion of UTRs was responsible for a total increase of approximately 0.002 of the genetic divergence compared with the data set including ORFs only (table 6). This corresponds to a relative increase of about 17%, 12.6%, and 8.7% in the genetic divergence between AB and OV and tilapia *versus* AB and OV, respectively. For the same gene data set, we also retrieved full-length contigs from AB and OV in order to extend our analysis of UTRs to longer sequences (thus excluding tilapia which reduced the length overlap across AB and OV in the previous data set). Average length of the 1,216 AB-OV pairwise alignments was 925 bp, nearly double

Table 5

Average Pairwise Genetic Distances (Pi, Uncorrected), Rates of Synonymous and Nonsynonymous Substitutions Per Site and Relative Ratio Estimated for Both Individual and Concatenated 1,216 Four-Species Alignments (526,113 bp, Data set #3)

			Individual Alignments			Concatenated Alignments					
			Nei and Gojobori (1986)			Nei and Gojobori (1986)			Goldman and Yang (1994)		
		Pi	Ks	Ka	Ka/Ks	Ks	Ka	Ka/Ks	d*S*	d*N*	d*N*/d*S*
AB	OV	0.0095 ± 0.0072	0.0289 ± 0.0001	0.0048 ± 4.7 × 10^{-06}	0.1856 ± 0.2688	0.0288	0.0057	0.1979	0.0288	0.0039	0.1358
Tilapia	AB	0.0222 ± 0.0207	0.0732 ± 0.0006	0.0096 ± 1 × 10^{-05}	0.1753 ± 0.2124	0.0685	0.0103	0.1504	0.0686	0.0091	0.1323
Tilapia	OV	0.0230 ± 0.0210	0.0746 ± 0.0005	0.0102 ± 0.0000	0.1827 ± 0.2349	0.0699	0.0117	0.1674	0.0700	0.0097	0.1387
Medaka	Tilapia	0.1609 ± 0.0496	0.8657 ± 0.0197	0.065 ± 0.0002	0.0810 ± 0.0977	0.8128	0.0672	0.0827	0.8160	0.0607	0.0744
Medaka	AB	0.1605 ± 0.0497	0.8695 ± 0.0125	0.0644 ± 0.0002	0.0806 ± 0.1171	0.8167	0.0665	0.0814	0.8201	0.06	0.0731
Meakda	OV	0.1605 ± 0.0497	0.8681 ± 0.016	0.0647 ± 0.0002	0.0810 ± 0.1062	0.8143	0.0676	0.0830	0.8182	0.0603	0.0737

Table 6

Average Pairwise Genetic Distances (Pi, Uncorrected) Estimated for 1,216 Individual Four-Species Alignments (Gene Data set #3) before and after Trimming UTRs

		Pi		
		ORFs only	ORFs + UTRs[a]	ORFs + UTRs[b]
AB	OV	0.0095 ± 0.0072	0.0112 ± 0.0077	0.0133 ± 0.0080
Tilapia	AB	0.0222 ± 0.0207	0.0250 ± 0.0171	na
Tilapia	OV	0.0230 ± 0.0210	0.0250 ± 0.0171	na

[a] Total length: 652,849 bp.

[b] Total length: 1,122,962 bp.

the length for the two species considering ORFs only, and ranged between 402 and 3,669 bp, for a total of 1,122,962 bp. Average pairwise divergence between the two species was 0.0133, corresponding to an increase of 40% of their genetic divergence with respect to alignments including ORFs only (table 6). This last value is likely an underestimate of UTR contribution to the genetic divergence between AB and OV; indeed, in some cases, these longer alignments also include additional coding regions that were trimmed in data set#3 because they did not fully overlap with tilapia sequences (which are, on average, shorter than our AB and OV contigs).

Rates of Evolution and Signature of Disruptive Selection in the Cichlid Lineage

We used ORFs from data set#3 to estimate rates of evolution within cichlids (table 5). Based on the mean pairwise estimates on single alignments, the smallest average Ks value was found for the AB/OV comparison (0.0289), followed by similarly low values between tilapia and both OV and AB (0.0732 and 0.0746) and between medaka and all other species comparisons (0.8657–0.8695). The average Ks values calculated from the concatenated alignment were comparable (table 5).

Within cichlids, the average pairwise Ka/Ks ratios across the three species were also similar (0.175–0.186) but at least two times higher than for all pairwise comparisons between medaka and the three cichlids (0.081) (Whitney–Mann test, $P < 0.001$), suggesting disruptive selection in the cichlid lineage. Estimates of Ka/Ks based on average individual and concatenated alignments using the Nei and Gojobori method were similar and comparable to the estimates obtained using the more sophisticated model of substitutions from Goldman and Yang (1994) implemented in Codeml (Yang 2007).

We further tested the hypothesis for differential selective forces among lineages by comparing several branch models implemented in Codeml (PAML). Among the models tested, the free-ratio model, allowing one d*N*/d*S* for each branch, was significantly better than both the one-ratio model, which assigned the same d*N*/d*S* value to all branches, and the two-ratio model, which assigned to the medaka lineage a d*N*/d*S* value that differed from all other branches (LRT, $P < 0.001$ in both comparisons). According to the free-ratio model, the branches within the cichlid clade evolved with at least as twice as large d*N*/d*S* (0.1269–0.1607) compared with the branch at the base of the clade (0.0687) (fig. 2). The branch leading to medaka also showed a d*N*/d*S* value similar to that of cichlids; however, individual values of d*N* and d*S* were extremely low ($<10^{-4}$), impeding a reliable estimate of the ratio.

Positively Selected Genes

We screened all individual 1,216 alignments for pairwise Ka/Ks values higher than one and obtained a set of 33 genes that are putatively under positive selection in at least one pairwise comparison (table 7). Individual inspection of these gene alignments ruled out possible misalignments or chimeric structures. All 33 genes showed Ka/Ks > 1 exclusively within cichlids comparisons: 14 genes between AB and OV, 13 between tilapia and either AB or OV, five in two pairwises and one gene for all three-cichlids pairwise comparisons. No genes showed values of Ka/Ks > 1 between medaka and any of the three cichlid species. This is compatible with the lower d*N*/d*S* value assigned to the branch leading to the cichlid clade (reported above).

To further confirm these findings, all 33 individual genes were tested for positive selection in the framework of a phylogeny using the branch free-ratio model in Codeml. d*N*/d*S* was larger than one in one or more lineages in all the 33 genes, supporting the above results.

Discussion

Coverage and Functional Annotation of the Two Transcriptomes

Our transcriptome-wide study provides the first high-throughput 454 sequencing data available for eastern African cichlids and the largest current EST data set for cichlids. With nearly 647,000 reads assembled in more than 46,000 contigs, this data set offers the very first extensive genetic resource for a member of the Ectodini tribe, *O. ventralis* (OV), for which current molecular data were limited to few mitochondrial and nuclear genes only (see, e.g., Clabaut et al. 2005; Salzburger et al. 2007; Koblmuller et al. 2008). It also largely integrates current EST data available for *A. burtoni* (AB) (Salzburger et al. 2008). Comparative analysis of the new EST data set generated for this species (49,311 contigs) with the one already available in GenBank (10,312 contigs) via BlastN searches (1×10^{-50}) indicates an overlap of 6,935 contigs between data sets. More than 70% of the hits showed a sequence identity between 98% and 100%, confirming the quality of our EST sequences and providing a further coverage for a subset of them. Overall, combining the two data sets, the ESTs generated in this study contributed to more than 70% of unique new sequences, greatly enlarging the current coverage of the transcriptome for AB.

Table 7

Genes Under Putative Positive Selection Based on Pairwise Ka/Ks Values > 1

Pairwise	Gene	Length, bp	Pi	Ks	Ka	Ka/Ks
Single						
AB versus OV	Aquaporin fa-chip	396	0.0202	0.0103	0.0273	2.650
	Succinate dehydrogenase	450	0.0178	0.0085	0.0214	2.518
	20-beta-hydroxysteroid dehydrogenase	501	0.0140	0.0084	0.0159	1.893
	26s proteasome nonatpase regulatory subunit 9	636	0.0173	0.0140	0.0227	1.621
	Muscle-type creatine kinase ckm1	438	0.0092	0.0098	0.0151	1.541
	Darmin protein	363	0.0083	0.0061	0.0090	1.475
	Serine hydrolase-like protein	489	0.0226	0.0180	0.0247	1.372
	Tetratricopeptide repeat protein 35	600	0.0034	0.0078	0.0107	1.372
	Transmembrane protein 16f	357	0.0114	0.0120	0.0148	1.233
	Dead (asp-glu-ala-asp) box polypeptide 56	537	0.0075	0.0080	0.0098	1.225
	Novel protein (zgc:100919)	384	0.0131	0.0116	0.0139	1.198
	loc733309 protein	363	0.0138	0.0128	0.0142	1.109
	Alpha-sialyltransferase st3gal v	345	0.0116	0.0111	0.0119	1.072
	Trypsinogen 2	540	0.0315	0.0311	0.0325	1.045
Tilapia versus OV	Beta-galactoside-binding lectin	378	0.0212	0.0119	0.0243	2.042
	Decaprenyl-diphosphate synthase subunit 2	348	0.0201	0.0120	0.0231	1.925
	Elastase 2-like protein	540	0.0225	0.0152	0.0253	1.664
	cdc42-interacting protein 4 homolog	306	0.0132	0.0157	0.0167	1.064
	Cytochrome c oxidase subunit 4 isoform mitochondrial precursor	516	0.0177	0.0178	0.0182	1.022
	Regulator of g-protein signaling 18	417	0.0240	0.0218	0.0283	1.298
	Serum paraoxonase arylesterase 2	435	0.0300	0.0305	0.0336	1.102
	hbaa_serqu ame: full = hemoglobin subunit alpha-a ame: full = hemoglobin alpha-a chain ame: full = alpha-a-globin	426	0.0423	0.0403	0.0445	1.104
	Suppression of tumorigenicity 14 (colon epithin)	477	0.0359	0.0266	0.0400	1.504
Tilapia versus AB	Signal sequence alpha	528	0.0076	0.0078	0.0126	1.615
	Nadh dehydrogenase 1 alpha subcomplex subunit mitochondrial precursor	330	0.0182	0.0135	0.0199	1.474
	mgc85594 protein	402	0.0150	0.0123	0.0159	1.293
	ca++ cardiac fast twitch 1 like	447	0.0201	0.0180	0.0212	1.178
Two						
Tilapia versus OV	Annexin a4	534	0.0356	0.0361	0.0366	1.014
Tilapia versus AB			0.0300	0.0279	0.0314	1.125
Tilapia versus OV	Lipid phosphate phosphohydrolase 2	258	0.0233	0.0149	0.0268	1.799
Tilapia versus AB			0.0233	0.0149	0.0268	1.799
AB versus OV	39s ribosomal protein mitochondrial precursor	318	0.0126	0.0138	0.0165	1.196
Tilapia versus AB			0.0189	0.0138	0.0207	1.500
AB versus OV	Ubiquinol-cytochrome c rieske iron-sulfur polypeptide 1	441	0.0136	0.0096	0.0150	1.563
Tilapia versus OV			0.0159	0.0096	0.0181	1.885
AB versus OV	Epithelial cadherin precursor	651	0.0691	0.0671	0.0742	1.106
Tilapia versus OV			0.0799	0.0807	0.0857	1.062
Three						
AB versus OV	Cell cycle control protein 50a	372	0.0162	0.0109	0.0218	2.000
Tilapia versus OV			0.0431	0.0218	0.0558	2.560
Tilapia versus AB			0.0457	0.0439	0.0483	1.100

Note.—Of the 33 genes, 27 were found with Ka/Ks > 1 only in single cichlid pairwises, five in two pairwises, and one in all three pairwise comparisons.

Based on comparison of the number of proteins predicted for closely related fishes with those identified in our two EST libraries, the transcriptomes generated for both AB and OV cover at least half of their total proteomes. Specifically, the number of protein-coding genes ranges from a minimum of 18,523 in the highly compact genome of *Takifugu* (Aparicio et al. 2002) to up to 24,147 in *D. rerio* (http://www.sanger.ac.uk/Projects/D_rerio/). Taking these two values as a reference range for the expected number of protein-coding genes, ESTs from AB cover between 52% and 67% of the total protein-coding genes diversity (with 8,684 predicted proteins, see table 2), whereas ESTs from OV cover between 47% and 61% (7,671 proteins). It is, however, important to consider that ESTs represent, in most cases, partial transcripts, with a typical 3′-UTR bias introduced during the sequencing process, and thus, the actual coverage obtained for a full proteome (total length of the cDNA sequences transcribed) of both species is likely lower.

Comparative Transcriptomics between AB and OV

Comparative analyses of the functional annotation of more than 10,000 EST contigs for both AB and OV showed highly similar transcriptomes between the two species, in terms of both types and relative frequencies of GO categories expressed. The ten most represented GO terms per category were typically the same for both species, with very similar relative and absolute frequencies (fig. 1). An analogous comparative transcriptome analysis was recently performed for two closely related Central America cichlids (Elmer et al. 2010) and also showed a comparable functional annotation of their transcriptomes, with similar coverage of expressed GO categories (both as types and frequencies) between species. These categories are however differently represented compared with our data set, suggesting quite divergent transcriptome features between Central American and eastern African cichlids, although further analyses are needed to explore these differences.

A large portion of the transcriptomes of AB and OV (64 and 75% of the contigs, respectively) could not be annotated or had no BlastX matches to the protein nr database, suggesting that these sequences might represent novel proteins, unique to cichlids, fast evolving genes or UTRs. Recent studies in humans indicate that large parts of transcriptomes are indeed noncoding, although this remains unclear in fishes (Cheng et al. 2005). Further identification of these contigs via BlastN searches in the nt database provided a significant match only for 9% of these contigs in both species, suggesting that the large majority of these sequences (either translated or untranslated) might indeed be cichlid-specific, as result, for instance, of accelerated sequence evolution. Among those contigs that returned a significant match in the nt database, roughly half of them matched to 13 unique hits (i.e., AccNos), represented solely by two gene categories, immune and patterning genes, both in AB and OV. The two species also paired in terms of relative frequencies of these most represented contigs, indicating similar high expression levels of these transcripts in AB and OV. Among other hits that were less represented in terms of number of contig per hit, several matched to genes that are known to play a crucial role in rapid species evolution, such as *bmP4*, *pax6*, and color genes. Overall, this suggests that genes implied in key features of (cichlid) species, such as body morphology, coloration, development, and immunity represent a variable portion of the cichlid transcriptome (i.e., genes under accelerated evolution) with respect to other species, as predicted based on their function in processes typically under strong natural selection. Nevertheless, these findings should be taken with caution as we cannot exclude a bias in the type of sequences available in the nt database for closely related species to AB and OV, which would also bias the results of the BlastN searches.

Genetic Diversity between AB and OV

The two new transcriptomes presented here show up to 0.0175 uncorrected genetic divergence based on >4,000 pairwise alignments of putatively orthologues ESTs identified through a best reciprocal hit approach. It should be noted that all the alignments included both annotated and nonannotated sequences. When only annotated sequences are considered (using data set #2), the genetic diversity drops to 0.0138 between OV and AB. Furthermore, when only ORFs are considered (data set #3), the genetic diversity drops to nearly half (0.0095), suggesting that noncoding regions and nonannotated coding genes, such as putative novel or fast evolving genes, contributed to at least half of the total transcriptome divergence. In particular, UTR regions appear to carry a great proportion of variable sites between the two species. Comparative analysis of the same gene data set before and after trimming UTRs indicates a 40% increase of genetic divergence between AB and OV when UTRs are included. Similarly, an increase in genetic divergence, although smaller (likely due to shorter sequences), is seen when UTRs are included in pariwise comparisons between tilapia and both AB and OV.

It has been proposed that large part of the phenotypic variation found among closely related species is associated to changes at the regulatory regions affecting the expression profiles (e.g., *cis*-regulatory elements; Fay and Wittkopp 2008). In cichlids, this scenario is mainly supported by the indirect finding of very limited or no genetic diversity at the protein-coding regions among phenotypically diverse species (see Kobayashi et al. [2009] for lake Victoria species and Elmer et al. [2010] for Central American cichlids). Direct evidence of adaptive variation at noncoding regions comes from recent data showing that cichlid 3′-UTRs contain target sites for fast evolving microRNA. These sites present elevated SNP densities in response to the rapid diversification of these miRNA, clearly pointing to a prominent role of UTRs in cichlid evolution (Loh et al. 2011).

In our data set, part of the observed UTR diversity might simply result from weaker functional constraints and therefore be nonadaptive. Future investigations targeting, for example, the functional role of divergent UTRs found in association with highly conserved protein-coding sequences will shed light on the contribution of UTRs in cichlids evolution.

Evolutionary Divergence and Mutational Rates among Cichlids

In order to address more specific questions on genetic diversity, substitution rates and selection within the cichlid clade, we expanded our comparative transcriptome analyses to include EST data publicly available for another cichlid, tilapia (Lee et al. 2010), which is a representative of a distinct and more ancestral cichlid lineage, as well as cDNA from medaka, which is presently the closest fully sequenced outgroup to cichlids (Steinke et al. 2006).

We were able to generate a total of 1,216 clusters of aligned sequences (up to 526 Kb) containing exclusively ORFs that fully overlapped across the four species (data set #3). The stringent criteria used for clustering, including cut off *e* values for Blast searches set to 1.0 x 10^{-50}, best reciprocal Blast hits and removal of sequence clusters with stop codons in cichlids (putatively pseudogenes or novel truncated genes) likely prevented inclusion of paralogous sequences, providing a reliable data set for molecular evolution analyses. Nevertheless, inference of orthology should be taken with caution as transcriptomes are partial and might not represent all sequences belonging to a gene family, causing reciprocal best BlastN hits between paralogous sequences. Although we can largely exclude clustering of cichlid paralogous sequences that are members of old gene families (formed before the cichlid radiation), we cannot rule out clustering of sequences derived from more recent lineage-specific duplications for which only one copy was present in individual species data sets.

Within cichlids, the nucleotide diversity Pi, Ka, and Ks between tilapia and both OV and AB was approximately the same but more than 2-fold higher than between OV and AB (table 5). This is also confirmed by the ML phylogeny reconstructed based on the concatenated data set, which shows equal branch length between tilapia and both AB and OV. Nucleotide diversity estimates based on nuclear data are available for other cichlids, too, albeit based on much smaller samples of orthologous genes. Specifically, genetic distances are reported for three members of the Lake Victoria region superflock, which range between 0.00339 and 0.00346 based on 68 genes (Kobayashi et al. 2009). An average genetic distance of 0.0026 was detected among five Malawi species, based on partial genomic data with low coverage (Loh et al. 2008).

Assuming a divergence time between tilapia and the remaining cichlids of 10.51 to 29.43 Ma (average of 19.44 Ma; Matschiner et al. 2011) and using the neutral Ks divergence estimated on the concatenated alignment by Codeml (accounting for transition/transversion rates and base-frequency dependency), we calculated a mutation rate ranging from 1.2 to 3.3 $\times$ 10^{-9} substitutions per silent site per year (average of 1.8 $\times$ 10^{-9} substitutions per silent site per year) for both comparisons of tilapia to OV and AB. This mutational rate is in accordance to the average mammalian genome mutation rate of 2.2 $\times$ 10^{-9} per base pair per year (Kumar and Subramanian 2002), but it could represent an underestimate because we did not correct for multiple hits. Using the linear equations of time *versus* Ks given by tilapia comparisons to AB and OV and considering a Ks value between AB and OV of 0.0288 (table 5), we estimated a divergence time for the AB-OV split of between 4.4 and 12 Ma (average of 8 Ma). This dating roughly coincides with the onset of truly lacustrine conditions in Lake Tanganyika (ca. 6 Ma), which is when the primary lacustrine radiation of cichlids is thought to have started and the main cichlid lineages, including the haplochromines and ectodines, emerged (see, e.g., Salzburger et al. 2002; Koblmuller et al. 2008).

Signature of Positive Selection in the Cichlid Lineage

Ka/Ks values estimated for all cichlid pairwise comparisons were at least two times greater (0.175–0.186) than those calculated between medaka and the three cichlids, which were nearly the same (0.081). This argues for homogeneous substitution rates within cichlids, independent of genetic divergence.

Looking at substitution rates in the framework of a phylogeny, d*N*/d*S* values per branch estimated under the best branch model (i.e., free-ratio model) confirmed a higher d*N*/d*S* for branches within the cichlid clade with respect to the outgroup medaka. This is largely concordant with previous findings for closely related Malawi cichlids, where cichlids showed a much higher Ka/Ks (up to five times) than the one estimated between more distant outgroups (such as between Fugu and *Tetraodon* or among *Danio* strains) (Loh et al. 2008). Taken as a whole, these studies provide good evidence for a relatively higher rate of fixation of nonsynonymous substitutions in cichlids, likely driven by disruptive selection. Alternatively, such elevated Ka/Ks might result, in part, from a relaxed purifying selection, due for instance to smaller effective population size of the cichlid ancestor.

Within our data set, we also specifically identified a set of 33 genes putatively under positive selection that represent potential candidates for a more thoroughly experimental and computational investigation. We note that the data set used for our estimates derived from randomly pooled ESTs that contained an ORF and showed a good level of amino acid conservation to return a significant BlastX hits to the medaka proteome, thus we do not expect any particular bias in the gene pooling. Nevertheless, these estimates should be taken with caution, as other types of biases should be considered. First, this data set comprises relatively short alignments of partial ORFs (mean was 433 bp), mostly due to a 3′-UTR bias introduced during the EST sequencing process. This decreases the power for testing positive selection in individual genes. Moreover, the cichlid radiation occurred in a very short evolutionary time frame and deleterious nonsynonymous mutations might not yet have been removed, which could affect proper estimates of Ka/Ks (Rocha et al. 2006; Wolf et al.

2009). Finally, 454 sequencing is known to have a high single read accuracy of > 99.5%, whereas the consensus read accuracy within an assembly with a coverage >20x is >99.99%. Nonetheless, even considering slightly higher error rates, we do not expect any systematic bias in the substitutions pattern that could have specifically affected the nonsynonymous rates.

Conclusions

Using the new 454 pyro-sequencing technology, we have provided the so far largest collection of new ESTs for cichlid species. Our functional annotation and expanded comparative transcriptome analysis, including a third cichlid lineage (tilapia) and the outgroup medaka, have shown a signature of disruptive selection in the cichlid lineage and pointed to a prominent contribution of UTRs in cichlid genetic diversity, potentially involved in regulatory changes of the expression profiles underlying their large phenotypic diversity. The new transcriptomes provide an important reference to now target more specific transcriptome-to-phenome comparative analyses aimed to investigate, for instance, the molecular bases of single and multiple traits diversity in more closely related species or shared traits among more distantly related species. Genome sequencing projects are currently ongoing for tilapia and four other cichlid species, including AB, *Metraclinia* (Maylandia) *zebra*, *Pundamilia nyererei*, and *Neolamprologus brichardi* (http://cichlid.umd.edu/CGCindex.html). Together with these, the partial genomic data and the EST resource already existing for cichlids and close outgroups, the transcriptome data sets reported here will provide the scientific community with a valuable resource for comparative analyses of both genetic and expression profiles within cichlids and among closely related species that will address crucial questions on the molecular bases of adaptive radiation and explosive speciation.

Supplementary Material

Supplementary table S1 is available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

We would like to thank Christof Wunderlin (Microsynth) for technical support on the 454 sequencing data and Michael Matschiner and Hugo Gante for helpful comments on the manuscript. This work was supported by a doctoral research fellowship from the Fundação para a Ciência e a Tecnologia [SFRH/BD/43421/2008 to E.S.], the European Research Council [Starting Grant 'INTERGENADAPT' to W.S.], and the Swiss National Science Foundation [grant 3100A0_122458 to W.S.].

Literature Cited

Aparicio S, et al. 2002. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science. 297:1301–1310.

Cheng J, et al. 2005. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science. 308:1149–1154.

Clabaut C, Salzburger W, Meyer A. 2005. Comparative phylogenetic analyses of the adaptive radiation of Lake Tanganyika cichlid fish: nuclear sequences are less homoplasious but also less informative than mitochondrial DNA. J Mol Evol. 61:666–681.

Conesa A, et al. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 21:3674–3676.

Cooper TF, Rozen DE, Lenski RE. 2003. Parallel changes in gene expression after 20,000 generations of evolution in Escherichia coli. Proc Natl Acad Sci U S A. 100:1072–1077.

Elmer KR, et al. 2010. Rapid evolution and selection inferred from the transcriptomes of sympatric crater lake cichlid fishes. Mol Ecol. 19(Suppl 1):197–211.

Ewing B, Hillier L, Wendl MC, Green P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 8:175–185.

Fay JC, Wittkopp PJ. 2008. Evaluating the role of natural selection in the evolution of gene regulation. Heredity. 100:191–199.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol. 11:725–736.

Johnson TC, et al. 1996. Late Pleistocene desiccation of Lake Victoria and rapid evolution of cichlid fishes. Science. 273:1091–1093.

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

Kobayashi N, Watanabe M, Horiike T, Kohara Y, Okada N. 2009. Extensive analysis of EST sequences reveals that all cichlid species in Lake Victoria share almost identical transcript sets. Gene. 441:187–191.

Koblmuller S, et al. 2008. Age and spread of the haplochromine cichlid fishes in Africa. Mol Phylogenet Evol. 49:153–169.

Kocher TD. 2004. Adaptive evolution and explosive speciation: the cichlid fish model. Nat Rev Genet. 5:288–298.

Kumar S, Subramanian S. 2002. Mutation rates in mammalian genomes. Proc Natl Acad Sci U S A. 99:803–808.

Lee BY, et al. 2010. An EST resource for tilapia based on 17 normalized libraries and assembly of 116,899 sequence tags. BMC Genomics. 11:278.

Loh YH, et al. 2008. Comparative analysis reveals signatures of differentiation amid genomic polymorphism in Lake Malawi cichlids. Genome Biol. 9:R113.

Loh YH, Yi SV, Streelman T. 2011. Evolution of microRNAs and the diversification of species. Genome Biol Evol. 3:55–65.

Matschiner M, Hanel R, Salzburger W. 2011. On the origin and trigger of the notothenioid adaptive radiation. PLoS One. 6:e18911.

Muller K. 2005. SeqState: primer design and sequence statistics for phylogenetic DNA datasets. Appl Bioinformatics. 4:65–69.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 3:418–426.

Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. 1999. The use of gene clusters to infer functional coupling. Proc Natl Acad Sci U S A. 96:2896–2901.

Rocha EP, et al. 2006. Comparisons of dN/dS are time dependent for closely related bacterial genomes. J Theor Biol. 239:226–235.

Salzburger W. 2009. The interaction of sexually and naturally selected traits in the adaptive radiations of cichlid fishes. Mol Ecol. 18:169–185.

Salzburger W, Braasch I, Meyer A. 2007. Adaptive sequence evolution in a color gene involved in the formation of the characteristic egg-dummies of male haplochromine cichlid fishes. BMC Biol. 5:51.

Salzburger W, Mack T, Verheyen E, Meyer A. 2005. Out of Tanganyika: genesis, explosive speciation, key-innovations and phylogeography of the haplochromine cichlid fishes. BMC Evol Biol. 5:17.

Salzburger W, Meyer A, Baric S, Verheyen E, Sturmbauer C. 2002. Phylogeny of the Lake Tanganyika cichlid species flock and its relationship to the Central and East African haplochromine cichlid fish faunas. Syst Biol. 51:113–135.

Salzburger W, et al. 2008. Annotation of expressed sequence tags for the East African cichlid fish Astatotilapia burtoni and evolutionary analyses of cichlid ORFs. BMC Genomics. 9:96.

Seehausen O. 2006. African cichlid fish: a model system in adaptive radiation research. Proc Biol Sci. 273:1987–1998.

Shapiro MD, et al. 2004. Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. Nature. 428:717–723.

Simmons MP, Ochoterena H. 2000. Gaps as characters in sequence-based phylogenetic analyses. Syst Biol. 49:369–381.

Stamatakis A, Ludwig T, Meier H. 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics. 21:456–463.

Steinke D, Salzburger W, Meyer A. 2006. Novel relationships among ten fish model species revealed based on a phylogenomic analysis using ESTs. J Mol Evol. 62:772–784.

Sturmbauer C, Meyer A. 1993. Mitochondrial phylogeny of the endemic mouthbrooding lineages of cichlid fishes from Lake Tanganyika in eastern Africa. Mol Biol Evol. 10:751–768.

Swofford DL. 2000. PAUP*: Phylogenetic Analysis Using Parsimony (*and other methods). Sunderland (MA): Sinauer Associates.

Verheyen E, Salzburger W, Snoeks J, Meyer A. 2003. Origin of the superflock of cichlid fishes from Lake Victoria, East Africa. Science. 300:325–329.

Wolf JB, Kunstner A, Nam K, Jakobsson M, Ellegren H. 2009. Nonlinear dynamics of nonsynonymous (dN) and synonymous (dS) substitution rates affects inference of selection. Genome Biol Evol. 1:308–319.

Wray GA, et al. 2003. The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol. 20:1377–1419.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.

Associate editor: Yves Van De Peer

Chapter 3

The evolution of cichlid egg-spots is linked with a *cis*-regulatory change

M. Emília Santos, Ingo Braasch, Nicolas Boileau, Britta S. Meyer, Loïc Sauteur, Astrid Boehne, Heinz-Georg Belting, Markus Affolter and Walter Salzburger

In press in Nature Communications

Nature Communications, manuscript NCOMMS-14-01786, final version

The evolution of cichlid fish egg-spots is linked with a *cis*-regulatory change

M. Emília Santos[1,2], Ingo Braasch[3], Nicolas Boileau[1], Britta S. Meyer[1], Loïc Sauteur[4], Astrid Böhne[1], Heinz-Georg Belting[4], Markus Affolter[4], and Walter Salzburger[1]

[1]Zoological Institute, University of Basel, 4051 Basel, Switzerland. [2]Present address: Institut de Génomique Fonctionnelle de Lyon, Ecole Normale Supérieure, CNRS UMR 5242, 46 Allée d'Italie, 69364 Lyon Cedex 07, France.[3]Institute of Neuroscience, University of Oregon, Eugene, OR 97403-1254. [4]Biozentrum, University of Basel, 4056 Basel, Switzerland. Correspondence and requests for materials should be addressed to M.E.S (email: emilia.p.santos@gmail.com) and/or to W.S. (email: walter.salzburger@unibas.ch).

Abstract

The origin of novel phenotypic characters is a key component in organismal diversification; yet, the mechanisms underlying the emergence of such evolutionary novelties are largely unknown. Here, we examine the origin of egg-spots, an evolutionary innovation of the most species-rich group of cichlids, the haplochromines, where these conspicuous male fin color markings are involved in mating. Applying a combination of RNAseq, comparative genomics and functional experiments, we identify two novel pigmentation genes, *fhl2a* and *fhl2b*, and show that especially the more rapidly evolving b-paralog is associated with egg-spot formation. We further find that egg-spot bearing haplochromines, but not other cichlids, feature a transposable element in the *cis*-regulatory region of *fhl2b*. Using transgenic zebrafish we finally demonstrate that this region shows specific enhancer activities in iridophores, a type of pigment cells found in egg-spots, suggesting that a *cis*-regulatory change is causally linked to the gain of expression in egg-spot bearing haplochromines.

[150 words]

The *de novo* evolution of complex phenotypic traits poses a challenge to evolutionary biology[1–5]. While selection explains adaptation and speciation in an adequate manner[6], it is more difficult to conceive how selection would trigger the origin of evolutionary novelties such as insect wings, feathers, the tetrapod limb, flowers, the mammalian placenta, beetle horns, or butterfly eye-spots[1,4,5,7,8]. The emergence of evolutionary innovations, *i.e.,* lineage restricted traits linked to qualitatively new functions, involves the origin of new developmental modules that are responsible for the identity of these novel characters[4,5]. Most of the available evidence suggests that new developmental programs emerge largely through co-option of pre-existing regulatory gene networks via changes in their regulation and deployment ("old genes playing new tricks"[5]). Uncovering the mechanisms of how these developmental modules are co-opted or newly evolved is one of the primary goals of evo-devo research[2,3,5,7,8].

Anal fin egg-spots are an evolutionary innovation in the so-called 'haplochromines'[9] (Fig. 1a; Supplementary Fig. 1), the most species-rich group of cichlid fishes best known for their spectacular adaptive radiations in the East African lakes Victoria and Malawi[10,11]. Adult males of approximately 1,500 cichlid species feature this pigmentation trait in the form of conspicuously colored circular markings[9,11,12]. Haplochromine egg-spots vary substantially in color, shape, number and arrangement between species (Fig. 1b), and even within species a certain degree of variation is observed. In some species, also females show egg-spots, which are then much less pronounced and colorful. The function of egg-spots has been implicated with the mating behavior of the female-mouthbrooding haplochromines[12,13]: Immediately upon spawning, a haplochromine female gathers up her eggs into the mouth; the male then presents its egg-spots to which the female responds by snatching, bringing her mouth close to the male's genital opening; upon discharging sperm, the eggs become fertilized inside the female's mouth (Fig. 1c). The mother subsequently broods and carries her progeny in the oral cavities for several weeks after fertilization.

Here, we are interested in the molecular basis of the anal fin egg-spots of haplochromine cichlids. The main advantages of the cichlid egg-spot system are that (*i*) the evolutionary innovation of interest emerged just a few million years ago and hence is recent compared to most other evolutionary novelties studied so far[9,10,14]; (*ii*) the phylogenetic context in which the novel trait evolved is known and living sister clades to the lineage featuring the novelty still exist[9,15,16]; and (*iii*) the genomes of two outgroup species lacking the trait and of three derived species featuring the trait are available. This allows us to study early events involved in the origin of an evolutionary innovation in an assemblage of phenotypically diverse, yet closely related and genetically similar species[14]. Using RNAseq, we identify two novel candidate pigmentation genes, the a- and b-paralogs of the four and a half LIM domain protein 2 (*fhl2*), and show that both genes, but especially the more rapidly evolving b-copy, are associated with the formation of egg-spots. We then find that egg-spot bearing haplochromines – but not an egg-spot-less ancestral haplochromine and not the representatives from more basal cichlid lineages – exhibit a transposable element insertion in close proximity to the transcription initiation site of *fhl2b*. A functional assay with transgenic zebrafish reveals that only a haplochromine-derived genetic construct featuring the SINE insertion drove expression in a special type of pigment cells, iridophores. Together, our data suggest that a *cis*-regulatory change (probably in the form of a SINE insertion) is responsible for the gain of expression of *fhl2b* in iridophores, contributing to the evolution of egg-spots in haplochromine cichlids.

Results

***fhl2a* and *fhl2b* as novel candidates for egg-spot morphogenesis.** As a first step, we performed an Illumina-based comparative transcriptomic experiment (RNAseq) between male (with egg-spots) and female (without egg-spots) anal fins in the haplochromine cichlid *Astatotilapia burtoni*. Two of the most differentially expressed genes according to RNAseq were the a- and b-paralogs of *fhl2* (~4 $\log_2$-fold and ~5 $\log_2$-fold differences, respectively; see Supplementary Table 2). These paralogs result from the teleost genome duplication[17] (Supplementary Fig. 2). The four and a half LIM domain protein 2 (Fhl2) is known as a transcriptional co-activator of the androgen receptor and the *Wnt*-signaling pathway[18,19]; Fhl2 plays a role in cell-fate determination and pattern formation; in the organization of the cytoskeleton, in cell adhesion, cell motility and signal transduction; furthermore, it regulates the development of heart, bone and musculature in vertebrates[20,21].

Expression of *fhl2a* and *fhl2b* is egg-spot specific and independent of patterning effects. To confirm the results obtained by RNAseq we performed quantitative real-time PCR experiments (Fig. 2a), this time also comparing egg-spot *versus* non-egg-spot tissue within male anal fins. In addition, we tested another haplochromine species with a different egg-spot arrangement to exclude positional effects of gene expression on the anal fin. In both species the two duplicates of *fhl2* were overexpressed in egg-spots (*A. burtoni*: *fhl2a*: t_5=10.77, p=0.0001; *fhl2b:* t_5=4.362, p=0.0073; *Cynotilapia pulpican*: *fhl2a*: t_4=5.031, p=0.0073; *fhl2b:* t_4=9.154, p=0.0008). We then tested the expression of both *fhl2* paralogs in the four main developmental stages of egg-spot formation in *A. burtoni*[22] and compared it to other candidate pigmentation genes (including the previously identified xanthophore marker *csf1ra*, the melanophore marker *mitfa*, and the iridophore marker *pnp4a*). We found that the expression of both *fhl2* paralogs increases substantially throughout anal fin and egg-spot development, and both genes show higher expression levels compared to the other pigmentation genes (Fig. 2b); *fhl2b* shows the highest increase in expression exactly when egg-spots begin to form. Furthermore, we corroborate that the expression domain of both *fhl2a* and *fhl2b* matches the conspicuously colored inner circle of egg-spots with RNA *in situ* hybridization (see Fig. 2c for results on *fhl2b*).

***fhl2a* and *fhl2b* evolved under purifying selection and show little polymorphism.** In general, phenotypic differences can arise via mutations affecting the function of proteins or via changes in gene regulation[5]. Therefore, we examined coding sequence evolution in the two *fhl2* paralogs to test for positive selection and potential change of function in a phylogenetically representative set of 26 East African cichlids. We found that the two *fhl2* genes are highly conserved in cichlids, with few amino acid differences between species and an average genetic divergence (0.4% in *fhl2a* and 0.7% in *fhl2b*) that lies below the transcriptome-wide average of 0.95%[23]. None of the observed amino acid changes was correlated with the egg-spot phenotype (Supplementary Table 7).

***fhl2b* shows greater functional specialization in haplochromines.** Usually, after a gene duplication event, the duplicates go through a period of relaxed selection, during which one of the two copies can diversify and acquire new functions[24]. We found that the b-copy of *fhl2* shows an elevated rate of molecular evolution compared to its paralog (*fhl2a*), which more closely resembles the ancestral sequence (Fig. 3a). An additional series of quantitative real-time PCR experiments in twelve tissues revealed that, in cichlids, *fhl2a* is primarily expressed in heart, bony structures and muscles, whereas *fhl2b* is highly expressed in the eye, and further in skin and the egg-spots of haplochromines (Fig. 3b and Fig. 3c). This is different to the gene expression profiles in medaka, where both duplicates are highly expressed in heart, skin and eye tissues; and in zebrafish, where the two paralogs are primarily expressed in heart, eye and (pharyngeal) jaw tissues, with *fhl2a* showing rather low levels of gene expression (Supplementary Figs. 3,4). When compared to the other teleost fishes examined here, our results suggest that the haplochromine *fhl2a* retained most of the previously described functions, whereas the more rapidly evolving *fhl2b* obtained new expression patterns. Together, the gene expression profile and the pattern of sequence evolution make this gene a prime candidate gene for the morphogenesis of haplochromine egg-spots.

***fhl2b* shows an AFC-SINE insertion only present in species with egg-spot.** Since there were no changes in the coding regions of *fhl2a* and *fhl2b* that are specific to the egg-spot-bearing haplochromines, we shifted our focus towards the analysis of putative regulatory elements, exploring the recently available genomes of five East African cichlids (including the egg-spot bearing haplochromines *A. burtoni*, *Pundamilia nyererei*, *Metriaclima zebra* and the egg-spot-less non-haplochromines *Neolamprologus brichardi* and *Oreochromis niloticus*). The non-coding region of *fhl2a* shows homology with other teleosts (*Oryzias latipes*, *Takifugu rubripes*, *Tetraodon nigroviridis* and *Gasterosteus aculeatus*) and we identified four conserved non-coding elements (CNEs) in all species examined (Supplementary Fig. 5a). These CNEs might thus represent conserved regulatory regions responsible for ancestral conserved functions of *fhl2a* in teleosts. We might be missing cichlid specific regulatory regions in important upstream regions though, as our capacity to detect lineage specific enhancers is limited due to the small sample size for each lineage and the high background conservation level present in cichlids.

Concerning *fhl2b* we did not find any CNE that is shared by cichlids and other teleosts (Supplementary Fig. 5b). Strikingly, however, we found a major difference that is shared by the three egg-spot bearing haplochromines: the presence of a transposable element upstream of *fhl2b*. Specifically, we identified a Short Interspersed Repetitive Element (SINE) belonging to the cichlid-specific AFC-SINEs (African cichlid family of SINEs[25]), which inserted ~800bp upstream of the transcriptional start site of *fhl2b* (Supplementary Fig. 6). To confirm that this insertion is associated with the egg-spot phenotype, we sequenced the upstream region of *fhl2b* in 19 cichlid species. The insertion was indeed present in nine additional, egg-spot bearing haplochromine species, yet absent in all ten non-haplochromines examined (Supplementary Table 8). Importantly, we found that one haplochromine species lacks the AFC-SINE element, namely *Pseudocrenilabrus philander.* This species belongs to one of the basal lineage of haplochromines (Fig. 1a) that is characterized by the absence of egg-spots (Fig. 1b). This suggests that the AFC-SINE upstream of *fhl2b* is not characteristic to the entire haplochromine clade, but to those that feature egg-spots, thus linking the SINE insertion to the origin of this evolutionary innovation.

The regulatory region of *fhl2b* from egg-spot-bearing haplochromines drives expression in iridophores in transgenic zebrafish. A long-standing hypothesis proposes that ubiquitous genomic repeat elements are potential regulators of transcription and could thereby generate evolutionary variations and novelties[26,27]. SINEs are known for their capability of 'transcriptional rewiring', i.e. to change the expression patterns of genes by bringing along new regulatory sequences, when inserted in close proximity to a gene's transcriptional initiation site[7,28]. In order to test whether the insertion of an AFC-SINE close to *fhl2b* functions as an enhancer of gene expression, we aimed for a functional experiment. We were particularly interested to find out whether there were changes in enhancer activity between AFC-SINE positive haplochromines and other cichlids lacking both the insertion and the egg-spot phenotype. To this end, we designed reporter constructs containing the upstream region of *fhl2b* (~2 kb upstream to intron 1) of three cichlid species linked to the coding region of Green Fluorescent Protein (GFP), and injected these constructs into zebrafish (*Danio rerio*) embryos to generate transgenic lines. We switched to the zebrafish system here, as no functioning transgenesis was available for haplochromine cichlids at the time the study was performed (due to the small number of eggs per clutch associated with the characteristic female mouthbrooding behavior). The three constructs were derived from *A. burtoni* (haplochromine with egg-spots, AFC-SINE$^+$), *P. philander* (haplochromine without egg-spots, AFC-SINE$^-$) and *Neolamprologus sexfasciatus* (lamprologine, AFC-SINE$^-$), respectively (Fig. 4a).

We were able to produce stable transgenic zebrafish lines for each of the three constructs to examine the expression of GFP. Importantly, we found striking differences in expression between the *A. burtoni* construct and the two constructs lacking the AFC-SINE. Of the three reporter lines only the AFC-SINE$^+$ showed GFP expression in iridophores, a silvery-reflective type of pigment cells (Fig. 4b,c and Supplementary Fig. 7). This experiment demonstrates the presence of novel enhancer activities in the regulatory region of *fhl2b* in derived haplochromines and strongly suggests that these came along with the SINE insertion.

Iridophores and egg-spot development. The egg-spot phenotype has previously been associated with pigment cells containing pteridines (xanthophores)[16,22], whereas our new results indicate an auxiliary role of iridophores in egg-spot formation. We thus re-evaluated the adult egg-spot phenotype by removing the pteridine pigments of the xanthophores (Fig. 4e). We indeed found that *A. burtoni* egg-spots show a high density of iridophores, which is further corroborated by the increase in gene expression of the iridophore marker *pnp4a* during egg-spot formation (Fig. 2b). With the exception of the proximal region of the anal fin, the number of iridophores is greatly reduced in the fin tissue surrounding egg-spots (Supplementary Fig. 8a). Interestingly, this proximal region is the only area of the anal fin besides the egg-spots where we observed *fhl2* expression with RNA *in situ* hybridization (see Fig. 2c for *fhl2b*), once more linking *fhl2* expression with iridophores (and less so with xanthophores, which are very rare in this region). In the non-haplochromine *N. crassus*, which features a yellow anal fin pattern containing xanthophores, we did not find iridophores in the xanthophore-rich region (Supplementary Fig. 9), suggesting that the xantophore/iridophore pattern is unique to haplochromine egg-spots.

Importantly, we also observed that iridophores appear early in the newly forming egg-spot of haplochromines, i.e. before the first xanthophores start to aggregate (Supplementary Fig. 8b).

In zebrafish, stripe development is initiated by iridophores, which serve as morphological landmarks for stripe orientation in that they attract further pigment cells such as xanthophores by expressing the *csf1* ligand gene[29,30]. Interestingly, it has previously been shown that a gene encoding a Csf1 receptor known for its role in xanthophore development in zebrafish, *csf1ra*, is expressed in haplochromine egg-spots[16]. We thus examined the expression of the ligand *csf1b* and show that its relative level of gene expression doubles during egg-spot development, and that this increase coincides with the emergence of the phenotype (Supplementary Fig. 10). This leads us to suggest that a similar pigment cell type interaction mechanism might be involved in egg-spot patterning as the one described for zebrafish[29,30]. The specific mode of action of fin patterning in haplochromine cichlids, and how Fhl2b interacts with the Csf1/Csf1r system, remains to be studied in the future.

Contribution of *fhl2a* in egg-spot formation. The role of the more conserved and functionally constrained a-paralog of *fhl2* in egg-spot development cannot be dismissed, though. Its temporally shifted increase in gene expression compared to *fhl2b* (Fig. 2b) suggests that *fhl2a* most likely acts as a more downstream factor involved in pigment pattern formation. We were nevertheless interested in uncovering the regulatory region responsive for this expression pattern. The first intron of *fhl2a* shows two conserved non-coding elements (CNEs) that are common across percomorph fish (Supplementary Fig. 5). Using the same strategy as described above we generated a transgenic zebrafish line containing exon 1 and intron 1 of *A. burtoni* linked to GFP. This construct drove expression in heart in zebrafish embryos, which is consistent with the reported function of *fhl2a* in tetrapods[20], whereas there was no indication of a pigment cell related function for this reporter construct (Supplementary Fig. 7e). An alignment between the genomic regions of the two *fhl2* paralogs shows that there were no CNEs in common and generally very little homology between them, suggesting that the regulation of the expression of *fhl2a* in egg-spots might proceed in a different way (Supplementary Fig. 11).

Discussion

In this study, we were interested in the genetic and developmental basis of egg-spots, an evolutionary innovation of the most species-rich group of cichlids, the haplochromines, where these conspicuous color markings on the anal fins of males play an important role in mating[11-13] (Fig. 1).

We first performed a comparative RNAseq experiment, which led to the identification of two novel candidate pigmentation genes, the a- and b-paralogs of the four and a half LIM domain protein 2 (*fhl2*). We then confirmed, with qPCR and RNA *in situ* hybridization, that the expression domain of both gene copies indeed matches the conspicuously colored inner circle of egg-spots (Fig. 2). Especially the more rapidly evolving b-copy of *fhl2* emerged as strong candidate gene for egg-spot development, as its expression profile mimics the formation of egg-spots (Figs. 2b, 3). Interestingly, we found that the egg-spot bearing haplochromines, but not other cichlids, feature a transposable element in the *cis*-regulatory region of *fhl2b*. Finally, we could show, making use of transgenic zebrafish, that a *cis*-regulatory change in *fhl2b* in the ancestor of the egg-spot bearing haplochromine cichlids (most likely in the form of the AFC-SINE insertion) resulted in a gain of expression in iridophores, a special type of pigment cells found in egg-spots (Fig. 4). This in turn might have led to changes in iridophore cell behavior and to novel interactions with pigmentation genes (*csf1b*, *csf1ra*, *pnp4a*), thereby contributing to the formation of egg-spots on male anal fins. The specific mode of action of the SINE insertion, and how the *fhl2b* locus interacts with these other pigmentation genes remains elusive at present. Addressing these questions would require functional studies in haplochromines, which are, however, hampered by the specific mechanisms involved in the trait complex of interest (mouthbrooding makes it notoriously difficult to obtain enough eggs – in a controlled manner – to make such experiments feasible).

Our results are also suggestive of an important role of the a-copy of *fhl2* in cichlid evolution. With our qPCR experiments, we provide strong evidence that *fhl2a* is involved in jaw tissue in zebrafish (Supplementary Fig. 3) and, importantly, in the pharyngeal jaw apparatus of cichlids (Fig. 3b,c), another putative evolutionary innovation of this group. The pharyngeal jaw apparatus is a second set of jaws in the pharynx of cichlids that is functionally decoupled from the oral jaws and primarily used to process food[11,12,15]. Interestingly, *fhl2a* has previously been implicated in the evolution of fleshy-lips in cichlids[31], which is yet another ecologically relevant trait. From a developmental perspective, the main tissues underlying these traits – the cranio-facial cartilage (the jaw apparatus) and pigment cells (egg- spots) – have the same origin, the neural crest, which itself is considered an evolutionary key innovation of vertebrates[32]. It thus seems that the function of *fhl2* in cichlids may have been split into (*i*) an ecologically important, *i.e.* naturally selected, scope of duties and (*ii*) a role in coloration and pigmentation more likely to be targeted by sexual selection.

Taken together, our study permits us to propose the following hypothesis for the origin of cichlid egg-spots: In one of the early, already female-mouthbrooding, haplochromines the insertion of a transposable element of the AFC-SINE family in the *cis*-regulatory region of *fhl2b*, and its associated recruitment to the iridophore pigment cell pathway, mediated the evolution of egg-spots on the anal fins – possibly from the so-called *perfleckmuster* common to many cichlids[16]. The conspicuous anal fin spots were fancied by haplochromine females, which – just like many other cichlids and also the ancestral and egg-spot less haplochromine genus *Pseudocrenilabrus* – have an innate bias for yellow/orange/red spots that resemble carotenoid-rich prey items[33], leading to the fixation of the novel trait. In today’s haplochromines, egg-spots seem to have a much broader range of functions related to sexual selection[34].

Most of the currently studied evolutionary innovations comprise relatively ancient traits (e.g., flowers, feathers, tetrapod limb, insect wings and mammalian placenta) making it difficult to scrutinize their genetic and developmental basis. Here, we explored a recently evolved novelty, the anal fin egg-spots of male haplochromine cichlids. We uncovered a regulatory change in close proximity to the transcriptional start site of a novel iridophore gene that likely contributes to the molecular basis of the origin of egg-spots in the most rapidly diversifying clade of vertebrates. This, once more, illustrates the importance of changes in *cis*-regulatory regions in morphological evolution[2].

Figure 1 | The egg-spots of haplochromine cichlids. (**a**) Phylogeny of the East African cichlid fishes based on a new multi-marker dataset. The haplochromines are the most species-rich and derived group of cichlids in East Africa. One of the common features of haplochromines is the presence of egg-spots on the anal fin of males. Note that one of the ancestral lineages, *Pseudocrenilabrus philander*, does not show this characteristic trait[9,33].

Substr-br: Substrate brooders; Mouthbr: Mouthbrooders; spp: species (**b**) Examples of male anal fin patterns in East African cichlids. Haplochromine egg-spots (upper panel) vary in size, shape, number and coloration (upper panel). Non-haplochromines and basal haplochromine *P. philander* (lower panel) do not show this trait (**c**) A typical mating cycle of haplochromine cichlids.

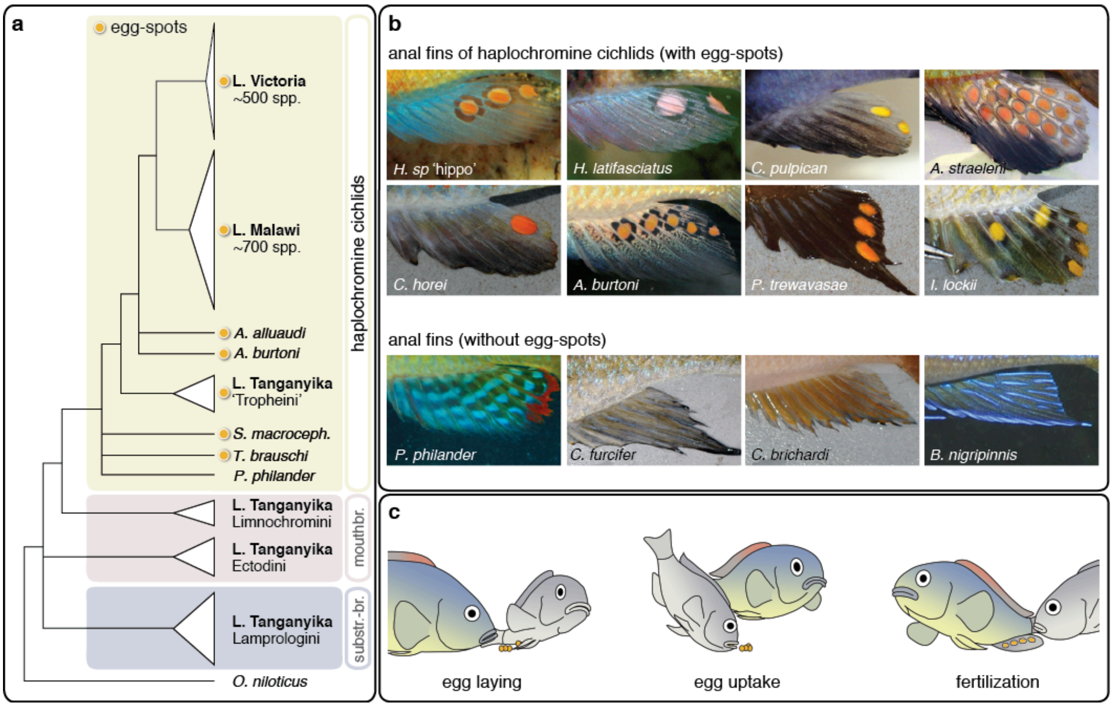


Figure 2 | The role of *fhl2a* and *fhl2b* in egg-spot formation. (**a**) Quantitative real-time PCR experiments reveal that both genes are overexpressed in egg-spot compared to adjacent anal fin tissue in the haplochromine cichlids *Astatotilapia burtoni* and *Cynotilapia pulpican* (* $p<0.05$; ** $p<0.01$; *** $p<0.001$; RQ = relative quantity). Images of males of the two species, their anal fins, and a scheme showing the distribution of egg-spots is provided. (**b**) Expression profiles of *fhl2a* and *fhl2b* during the ontogenetic development of egg-spots in *A. burtoni* (note that egg-spots are absent in juveniles and only form when males become sexually mature; see ref. 22 for further details). The values on the x-axis represent fish standard length in millimeters (three replicates per developmental stage were used). The error bars represent the standard error of the mean (SEM). *fhl2b* shows the largest increase in expression overall and its expression profile mimics the formation of egg-spots. Three other pigmentation genes (*pnp4a, csf1ra* and *mitfa*) were included for comparative reasons. *csf1ra* and *mitfa* show a much smaller increase in gene expression during egg-spot development than *fhl2a* and especially *fhl2b*, whilst *pnp4a* shows a constant increase in gene expression throughout the development of egg-spots. (**c**) RNA *in situ* hybridization experiments revealed that both *fhl2* paralogs (results only shown for *fhl2b*) are primarily expressed in the colorful inner circle of haplochromine egg-spots (defined by the solid line) and not in the transparent outer ring (defined by the dashed line). Expression was also observed in the proximal fin region, which also contains pigment cells. Panel 2 is a close up from the region defined by the square in panel 1.

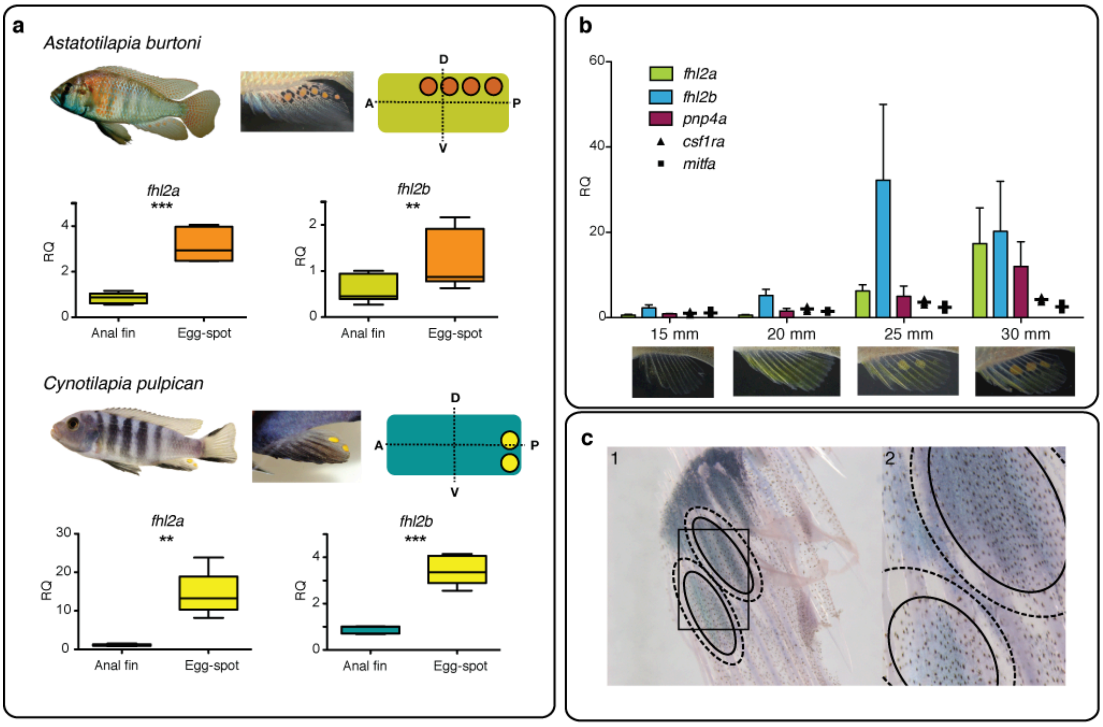


Figure 3 | Gene tree of the two *fhl2* paralogs and expression profiling in East African cichlid fishes. (**a**) Bayesian inference phylogeny of the orthology and paralogy relationships between cichlids, other teleosts (*Oryzias latipes*, *Danio rerio, Takifugu rubripes and Gasterosteus aculeatus*) and tetrapods (*Anolis carolinensis* and *Mus musculus*) *fhl2* sequences. This gene tree is important for generating functional hypotheses about both duplicates, and to infer the ancestral state of the *fhl2* gene before duplication. Our phylogeny indicates that *fhl2a* is more similar to the ancestral state, while *fhl2b* is apparently evolving faster in teleosts. Values at the tree nodes represent posterior probabilities. In Supplementary Fig. 2 we present a synteny analysis supporting the origin of teleost *fhl*2 duplicates in the teleost genome duplication. (**b**) Relative quantification (RQ) of *fhl2a* and *fhl2b* gene expression in twelve tissues (three replicates per tissue) in *C. pulpican*, an egg-spot bearing haplochromine from Lake Malawi. The error bars represent the standard error of the mean (SEM). (**c**) RQ of *fhl2a* and *fhl2b* gene expression in twelve tissues in *Neolamprologus crassus*, a substrate spawning lamprologine that has no egg-spots. In both species, gill tissue was used as reference; in *N. crassus*, "egg-spots" corresponds to the fin region where haplochromines would show the egg-spot trait. In *C. pulpican* (**b**), *fhl2a* is highly expressed in heart, in pigmented tissues (eye, skin and egg-spot) and in craniofacial traits (oral jaw and lower pharyngeal jaw); *fhl2b* is mainly expressed in the pigmented tissues. *N. crassus* (**c**) shows a similar expression patterns for *fhl2a* and *fhl2b*, with the difference that *fhl2a* does not show high expression levels in jaw tissues and *fhl2b* was not highly expressed in skin and fin tissue. These results suggest that *fhl2b* shows a higher functional specialization and that it might be involved in the morphogenesis of sexually dimorphic traits such as pigmented traits including egg-spots.

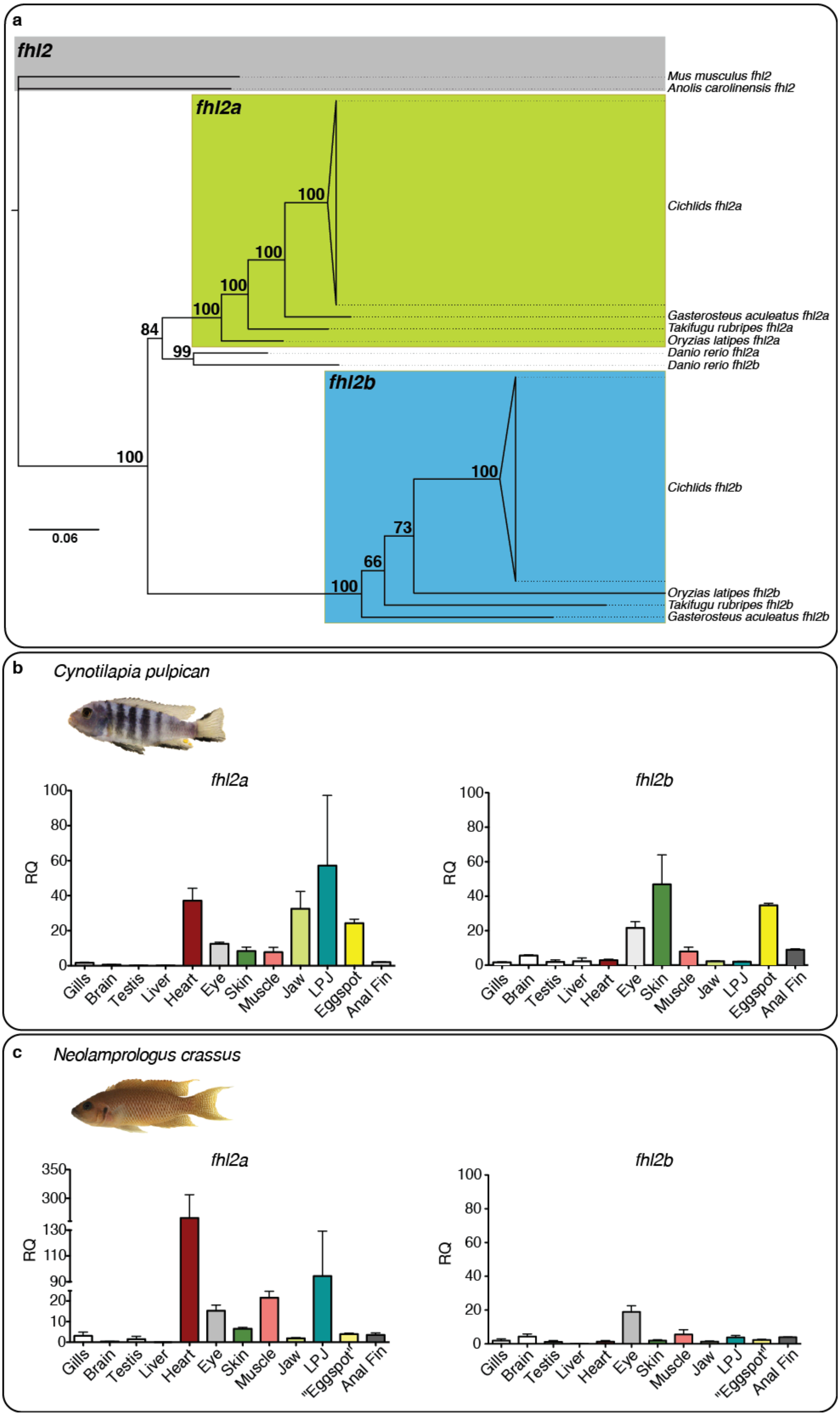
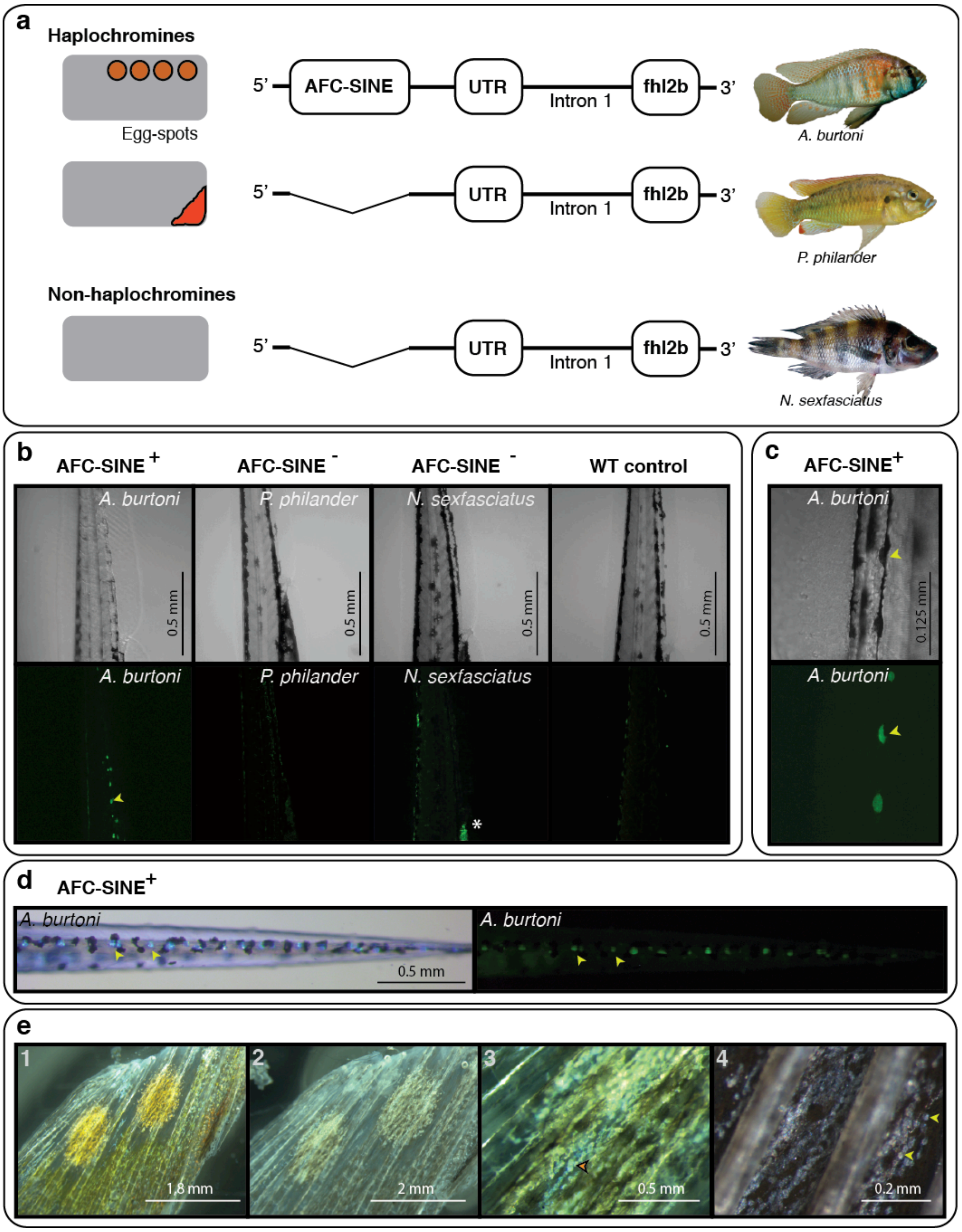

a
fhl2
Mus musculus fhl2
Anolis carolinensis fhl2
fhl2a
100
100
100
100
Cichlids fhl2a
Gasterosteus aculeatus fhl2a
Takifugu rubripes fhl2a
Oryzias latipes fhl2a
84
99
Danio rerio fhl2a
Danio rerio fhl2b
fhl2b
100
100
Cichlids fhl2b
0.06
73
66
100
Oryzias latipes fhl2b
Takifugu rubripes fhl2b
Gasterosteus aculeatus fhl2b
b
Cynotilapia pulpican
fhl2a
RQ
100
80
60
40
20
0
Gills
Brain
Testis
Liver
Heart
Eye
Skin
Muscle
Jaw
LPJ
Eggspot
Anal Fin
fhl2b
RQ
100
80
60
40
20
0
Gills
Brain
Testis
Liver
Heart
Eye
Skin
Muscle
Jaw
LPJ
Eggspot
Anal Fin
c
Neolamprologus crassus
fhl2a
RQ
350
300
130
110
90
20
10
0
Gills
Brain
Testis
Liver
Heart
Eye
Skin
Muscle
Jaw
LPJ
"Eggspot"
Anal Fin
fhl2b
RQ
100
80
60
40
20
0
Gills
Brain
Testis
Liver
Heart
Eye
Skin
Muscle
Jaw
LPJ
"Eggspot"
Anal Fin

Figure 4 | The molecular basis of egg-spot formation. (**a**) The egg-spot bearing haplochromines feature an AFC-SINE insertion in close proximity to the transcriptional start site of *fhl2b*, which is absent in the ancestral and egg-spot-less genus *Pseudocrenilabrus* and in all non-haplochromines. The sequences from the three species shown here were the ones used to engineer the reporter constructs, where the *fhl2b* coding sequence was substituted by GFP. (**b**) In transgenic zebrafish only the AFC-SINE$^+$ construct showed GFP expression in the iridophores, a type of pigment cell (one of them is indicated by a yellow arrow). The upper panel depicts bright field images of 3 day-old zebrafish embryo trunks; the lower panel shows the respective embryos under UV light. The green signal in the AFC-SINE negative *N. sexfasciatus* line (marked with an asterisk) is auto-fluorescence from the yolk extension. (**c**) Higher magnification image from *A. burtoni* AFC-SINE$^+$ reporter construct driving GFP expression in the iridophores. Orientation in (**b**) and (**c**): bottom: anterior, top: posterior. (**d**) Top-down view of a trunk of a 3 day-old AFC-SINE positive zebrafish embryo. The left panel depicts a bright field image where the iridophores of the dorsal stripe are illuminated by the incident light (yellow arrows). The right panel depicts GFP expression of the same embryo. The GFP signal co-localizes with iridophores. (**e**) Cellular basis of egg-spots: This series of images shows that egg-spots are made up of xanthophores, iridophores, and scattered melanophores. Image 1 shows an *A. burtoni* fin with two egg-spots. Image 2 shows the same fin without pteridine pigments (xanthophores are not visible anymore). Image 3 and 4 are higher magnification images of the egg-spots without pteridine under slightly different light conditions confirming that egg-spots have a high density of iridophores (examples of this cell type are highlighted with arrows).


a
Haplochromines
Egg-spots
5'
AFC-SINE
UTR
Intron 1
fhl2b
3'
A. burtoni
5'
UTR
Intron 1
fhl2b
3'
P. philander
Non-haplochromines
5'
UTR
Intron 1
fhl2b
3'
N. sexfasciatus
b
AFC-SINE +
AFC-SINE -
AFC-SINE -
WT control
A. burtoni
P. philander
N. sexfasciatus
0.5 mm
0.5 mm
0.5 mm
0.5 mm
A. burtoni
P. philander
N. sexfasciatus
*
c
AFC-SINE+
A. burtoni
0.125 mm
A. burtoni
d
AFC-SINE+
A. burtoni
0.5 mm
A. burtoni
e
1
1.8 mm
2
2 mm
3
0.5 mm
4
0.2 mm

Methods

Samples

Laboratory strains of *Astatotilapia burtoni*, *Cynotilapia pulpican*, *Astatoreochromis alluaudi, Pundamilia nyererei, Labidochromis caeruleus, Pseudotropheus elegans* and *Neolamprologus crassus* were kept at the University of Basel (Switzerland) under standard conditions (12h light/12h dark; 26 °C, pH 7). Before dissection, all specimens were euthanized with MS 222 (Sigma-Aldrich, USA) following an approved procedure (permit nr. 2317 issued by the cantonal veterinary office Basel). Individuals of all other specimens were collected in the southern region of Lake Tanganyika (Zambia) under the permission of the Lake Tanganyika Unit, Department of Fisheries, Republic of Zambia, and processed in the field following our standard operating procedure[15]. Tissues for RNA extraction were stored in RNAlater (Ambion, USA), tissues for genomic DNA extraction were stored in ethanol and shipped to the University of Basel.

RNA and DNA Extractions

Isolation of RNA was performed according to the TRIzol® protocol (Invitrogen, USA) after incubating the dissected tissues in 750µl of TRIzol at 4 ºC overnight or, alternatively, for 8-16 hours (in order to increase the RNA yield after long term storage). The tissues were then homogenized with a BeadBeater (FastPrep-24; MP, Biomedicals, France). Subsequent DNase treatment was performed with DNA-Free™ kit (Ambion, USA). RNA quantity and quality was determined with a NanoDrop 1000 spectrophotometer (Thermo Scientific, USA). cDNA was produced using the High Capacity RNA-to-cDNA kit (Applied Biosystems, USA). Genomic DNA was extracted using a high salt extraction method (modified from ref. 35).

Phylogenetic Analyses

DNA extraction of 18 specimens of East African cichlid fishes was conducted as described above. For the amplification of nine nuclear markers (*rag*, *gapdhs*, *s7*, *bmp4*, *ednrb1*, *mitfa*, *tyr*, *hag* and *csfr1*) we used the primer sets published in ref. 36. The sequences of *Metriaclima zebra, Oreochromis niloticus* and *Neolamprologus brichardi* were extracted from the respective genome assemblies (http://www.broadinstitute.org/models/tilapia). The data for *Astatoreochromis alluaudi, Thoracochromis brauschi* and *Serranochromis macrocephalus* were collected with Sanger sequencing following the method described in ref. 36, all other data were generated by amplicon sequencing with 454 GS FLX system at Microsynth, Switzerland, following the manufacturers protocols[37,38]. Sequences were quality filtered using PRINSEQ (length: 150bp minimum; low quality: mean $\geq$ 15; read duplicates)[39] and assembled with BWA-SW followed by visual inspection and consensus sequence generation in Geneious® 6.1.6[40]. As a tenth marker, we included mitochondrial NADH Dehydrogenase Subunit 2 (ND2) sequences available on GenBank (see Supplementary Table 1 for accession numbers). Since the *ednrb1* gene sequence is not available in the *N. brichardi* genome assembly, we used the gene sequence from its sister species, *N. pulcher*, instead.

Sequences were aligned with MAFFT[41] and the most appropriate substitution model of molecular evolution for each marker was determined with JMODELTEST v2.1.3[42] and BIC[43]. The partitioned dataset (5051bp) was then subjected to phylogenetic analyses in MRBAYES v3.2.1[44] and GARLI v2.0[45]. MRBAYES was run for 10,000,000 generations with 2 runs and 4 chains in parallel and a burn-in of 25%, GARLI was running 50 times followed by a bootstrap analysis with 500 replicates. SUMTREES v3.3.1 of the DENDROPY package v3.12.0[46] was used to summarize over the replicates.

Differential Gene Expression Analysis Using RNAseq

We used a transcriptomic approach (RNAseq) to identify genes differentially expressed between male and female anal fins of *Astatotilapia burtoni*. Library construction and sequencing of RNA extracted from three male and three female anal fins (at the developmental stage of 30 mm, see Fig. 2) was performed at the Department of Biosystems Science and Engineering (D-BSSE), University of Basel and ETH Zurich. The samples were sequenced on an Illumina Genome Analyzer IIx. Each sample was sequenced in one lane and with a read length of 76bp.

The reads were then aligned to an embryonic *A. burtoni* reference transcriptome assembled by Broad Institute (http://www.broadinstitute.org/models/tilapia). This transcriptome is not annotated and each transcript has a nomenclature where the first term codes for the parent contig and the third term codes for alternatively spliced transcripts (CompX_cX_seqX). The reference transcriptome was indexed using NOVOINDEX (www.novocraft.com) with default parameters. Using NOVOALIGN (www.novocraft.com), the RNAseq reads were mapped against the reference transcriptome with a maximum alignment (t) score of 30, a minimum of good quality base pair per read (l) of 25 and a successive trimming factor (s) of 5. Reads that did not match these criteria were discarded. Since the reference transcriptome has multiple transcripts/isoforms belonging to the same gene, all read alignment locations were reported (rALL). The mapping results were reported (o) in SAM format. The output SAM file was then transformed into BAM format, sorted, indexed and converted to count files (number of reads per transcript) using SAMTOOLS version 0.1.18[47]. The count files were subsequently concatenated into a single dataset - count table - and analyzed with the R package EDGER[48] in order to test for significant differences in gene expression between male and female anal fins. The ten most differentially expressed transcripts were identified by BLASTx[49] against GenBank's non-redundant database (see Supplementary Table 2).

We selected two genes out of this list for in-depth analyses – *fhl2a* and *fhl2b* – for three reasons: (*i*) *fhl2b* was the gene showing the highest difference in expression between male and female anal fins; (*ii*) The difference in gene expression in its paralog, *fhl2a*, was also significantly high; and (*iii*) the functional repertoire of the FHL2 protein family indicates that these might be strong candidates for the morphogenesis of a secondary male color trait.

Differential Gene Expression Analysis Using qPCR

The expression patterns of *fhl2a* and *fhl2b* were further characterized by means of quantitative real time PCR (qPCR) in three species, *A. burtoni*, *C. pulpican* and *N. crassus*. The comparative CT (cycle threshold) method[50] was used to calculate differences in expression between the different samples using the ribosomal protein L7 (*rpl7*) and the ribosomal protein SA3 (*rpsa3*) as endogenous controls. All reactions had a final cDNA concentration of 1ng/µl and a primer concentration of 200 mM. The reactions were run on a StepOnePlusTM Real-Time PCR system (Applied Biosystems, USA) using the SYBR Green master mix (Roche, Switzerland) with an annealing temperature of 58°C and following the manufacturers protocols. Primers were designed with the software GENSCRIPT REAL-TIME PCR (TAQMAN) PRIMER DESIGN available at https://www.genscript.com/ssl-bin/app/primer. All primers were designed to span over exons to avoid gDNA contamination (see Supplementary Table 3 for details). Primer efficiencies of the experimental primers (*fhl2a* and *fhl2b*) were comparable to the efficiency of the endogenous controls *rpl7* and *rpsa3*.

We conducted the following experiments: qPCR Experiment 1: Egg-spots were separated from the anal fin tissue in six male *A. burtoni* and five male *C. pulpican*. RQ (relative quantities) values were calculated for each sample and the differential expression between anal fin (reference) and egg-spot tissue was analyzed with a paired *t*-test using GRAPHPAD Prism version 5.0a for Mac OS X (www.graphpad.com). qPCR Experiment 2: *fhl2a, fhl2b, csf1ra, mitfa, pnp4a* and *csf1b* expression was measured in RNA extracted from *A. burtoni* fins at four different developmental stages[22]. Here, *csf1ra* was included as xanthophore marker[16], *mitfa* and *pnp4a* as melanophore and iridophore markers[51], respectively, and *csf1b* because of its role in pigment pattern organization in zebrafish[29,30]. We used three biological replicates for each developmental stage and each replicate consisted of a sample pool of three fins, except for the youngest stage at 15mm, where we pooled five fins. The first developmental stage was used as reference tissue. qPCR Experiment 3: *fhl2a* and *fhl2b* expression was measured in RNA extracted from different tissues from three males from *C. pulpican* and *N. crassus* (gills, liver, testis, brain, heart, eye, skin, muscle, oral jaw, pharyngeal jaw and egg-spot). Although *N. crassus* does not have egg-spots, we separated its anal fin into an area corresponding to egg-spots in haplochromines and a section corresponding to anal fin tissue (the ‘egg-spot’ region was defined according to the egg-spot positioning in *A. burtoni*). Expression was compared among tissues for each species using gills as reference tissue. The same experiment was performed for *D. rerio* and *O. latipes* (two teleost outgroups), using *ef1a* and *rpl13a*[52] as well as *rpl7* and *18sRNA*[53] as endogenous controls, respectively.

Cloning of *fhl2a* and *fhl2b* and RNA *in situ* Hybridization Experiments

Astatotilapia burtoni fhl2a and *fhl2b* coding fragments were amplified by PCR (for primer information see Supplementary Table 3) using Phusion® Master Mix with HF Buffer (New England BioLabs, USA) following the manufacturer’s guidelines. These fragments were cloned into pCR4-TOPO TA vector using the TOPO® TA cloning kit (Invitrogen, USA). Plasmid extractions were done with GenElute™ Plasmid Miniprep Kit (Sigma-Aldrich, USA). RNA probes were synthetized with the DIG RNA labeling kit (SP6/T7) (Roche, Switzerland). The insertion and direction of the fragments was confirmed by Sanger sequencing using M13 primers (available with the cloning kit) and BigDye® terminator reaction chemistry (Applied Biosystems, USA) on an AB3130*xl* genetic analyzer (Applied Biosystems, USA). *In situ* hybridization was performed in 12 fins from *A. burtoni* males, six for *fhl2a* and six for *fhl2b*. The protocol was executed as described in ref. 16, except for an intermediate proteinase *K* treatment (20 minutes at a final concentration of 15 µg/ml) and for the hybridization temperature (65ºC).

Synteny Analysis of Teleost *fhl2* Paralogs

The Synteny Database (http://syntenydb.uoregon.edu[54]) was used to generate dotplots of the human *FHL2* gene (ENSG00000115641) region on chromosome Hsa2 and the genomes of medaka (Supplementary Fig. 2a) and zebrafish (Supplementary Fig. 2b).

Double conserved synteny between the human *FHL2* gene and the *fhl2a* and *fhl2b* paralogons in teleost genomes provide evidence that the teleost *fhl2* paralogs were generated during the teleost genome duplication.

Sequencing of *fhl2a/fhl2b* Coding Region and Sequence Analysis

We then used cDNA pools extracted from anal fin tissue to amplify and sequence the coding region of *fhl2a* and *fhl2b* in a phylogenetically representative set of 26 cichlid species (21 Tanganyikan species, three species from Lake Malawi, and two species from the Lake Victoria basin). This taxon sampling included 14 species belonging to the haplochromines and 12 species belonging to other East African cichlid tribes not featuring the egg-spot trait (see Supplementary Table 4). *fhl2a* and *fhl2b* coding regions were fully sequenced (from start to stop codon) in five individuals per species in order to evaluate the rate of molecular evolution among cichlids. For PCR amplification, we used Phusion® Master Mix and cichlid specific primers (for primer information see Supplementary Table 3) designed with Primer3[55]. PCR products were visualized with electrophoresis in a 1.5% agarose gel using GelRed (Biotium, USA). In cases where multiple bands were present, we purified the correct size fragment from the gel using the GenElute™ Gel Extraction Kit (Sigma-Aldrich, USA). PCR products were enzymatically cleaned with ExoSAP-IT® (Affymetrix, USA) and sequenced with BigDye™ 3.1 Ready reaction mix (Applied Biosystems, USA) – after BigDye™ XTerminator purificaton (Applied Biosystems, USA) – on an AbI3130*xl* Genetic Analyzer. Sequences were corrected, trimmed and aligned manually in CODONCODE® ALIGNER (CodonCode Corporation).

fhl2 Phylogenetic Analysis

fhl2a and *fhl2b* sequences from non-cichlid teleosts and *fhl2* sequences from tetrapods were retrieved from ENSEMBL[56] (species names, gene names and accession numbers are available in Supplementary Table 5). We then constructed gene trees based on these sequences and on a subset of the cichlid sequences obtained in the previous step (information available in Supplementary Table 4) in order to confirm the orthologous and paralogous relationships of both duplicates. Sequences were aligned with CLUSTALW2[57] using default parameters. The most appropriate model of sequence evolution was determined with JMODELTEST as described above. Phylogenetic analyses were performed with MRBAYES (1 million generations; 25% burnin).

Tests for Positive Selection in *fhl2a* and *fhl2b*

Using PAUP* 4.0b10[58] we first compiled a maximum likelihood tree based on the mitochondrial ND2 gene, including all species used for the positive selection analyses (see Supplementary Table 6 for species and GenBank accession numbers). We used the GTR + Γ model with base frequencies and substitution rate matrix estimated from the data (as suggested by JMODELTEST[42]). We then ran CODEML implemented in PAML version 4.4b to test for branch-specific adaptive evolution in *fhl2a* and *fhl2b* applying the branch-site model (free-ratios model with ω allowed to vary)[59,60]. The branch comparisons and results are shown in Supplementary Table 7.

Identification of Conserved Non-Coding Elements

We then made use of the five available cichlid genomes[61] to identify conserved non-coding regions (CNEs) that could explain the difference in expression of *fhl2a* and *fhl2b* between haplochromines and non-haplochromines (note that there are three haplochromine genomes available: *A. burtoni, Pundamilia nyererei, Metriaclima* *zebra*; and two genomes belonging to more ancestral cichlid lineages: *Neolamprologus brichardi* and *Oreochromis niloticus*). For this analysis, we also included the respective genomic regions of four other teleost species (*Oryzias latipes*, *Takifugu rubripes*, *Tetraodon nigroviridis* and *Gasterosteus aculeatus*). More specifically, we extracted the genomic scaffolds containing *fhl2a* and *fhl2b* from the available cichlid genomes using BLAST v. 2.2.25 and the BIOCONDUCTOR R package BIOSTRINGS[62] to extract 5-6 kb of sequence containing *fhl2a* and *fhl2b* from these scaffolds.

Comparative analyses of the *fhl2a* and *fhl2b* genomic regions were done with MVISTA (genome.lbl.gov/vista)[63] using the LAGAN alignment tool[64]; *A. burtoni* was used as a reference for the alignment. We applied the repeat masking option with *Takifugu rubripes* (Fugu) as reference. Conserved non-coding regions (CNEs) were defined as any non-coding section longer than 100bp that showed at least 70% sequence identity with *A. burtoni*.

Sequencing of the Upstream Region of *fhl2b*

In order to confirm if the AFC-SINE insertion was specific to egg-spot bearing haplochromines, we amplified the genomic region upstream of the *fhl2b* open reading frame in 19 additional cichlid species (10 haplochromines and 9 non-haplochromines). PCR amplification was performed as described above. For sequencing, we used four different primers, the two used in the amplification reaction and two internal primers, one haplochromine specific and another non-haplochromine specific. For detailed information about species and primers see Supplementary Table 8.

Alignment of AFC-SINES from the *A. burtoni* Genome

SINE elements were identified using the SINE insertion sequence 5' of the *fhl2b* gene of *A. burtoni* as query in a local BLASTn search[49] with default settings against the *A. burtoni* reference genome. Blast hits were retrieved using custom scripts and extended to a region of 200bp upstream and downstream of the identified sequence. Sequences were aligned using MAFFT v. 6[41] with default settings and allowing for adjustment of sequence direction according to the reference sequence. The alignment was loaded into CODONCODE® ALIGNER for manual correction and end trimming. Sequences shorter than 50bp were excluded from the alignment. The final alignment contained 407 sequences that were used to build the *A. burtoni* SINE consensus sequence using the consensus method implemented in CODONCODE® ALIGNER with a percentage-based consensus and a cut-off of 25%. The AFC-SINE element in the *fhl2b* promoter region was compared to the consensus sequence and available full-length AFC-SINE elements of cichlids in order to determine whether it was an insertion or deletion in haplochromines (Supplementary Table 8).

Detailed Characterization of *fhl2b* Upstream Genomic Region in Cichlids

The *fhl2b* genomic regions of the five cichlid genomes (*A. burtoni*, *M. zebra*, *P. nyererei*, *N. brichardi*, and *O. niloticus*) were loaded into CODONCODE ALIGNER and assembled (large gap alignments settings, identity cut off 70%). Assemblies were manually corrected. Transposable element sequences were identified using the Repeat Masking function of REPBASE UNIT (http://www.girinst.org/censor/index.php) against all sequence sources and the bl2seq function of BLASTn[49]. Supplementary Fig. 6 shows a scheme of the transposable element composition of this genomic region in several cichlid species.

CNEs Construct Cloning and Injection in Zebrafish

We designed three genetic constructs containing the AFC-SINE and intron 1 of *fhl2b* of three cichlid species (*A. burtoni*, *P. philander* and *Neolamprologus sexfasciatus*) (Fig. 4) and one containing the 5'UTR, exon 1 and intron 1 of *A. burtoni fhl2a*. The three fragments were amplified with polymerase chain reaction as described above (see Supplementary Table 3 for primer information). All fragments were cloned into a pCR8/GW/TOPO vector (Invitrogen, USA) following the manufacturers specifications. Sequence identity and direction of fragment insertion were confirmed via Sanger sequencing (as described above) using M13 primers. All plasmid extractions were performed with GenEluteTM Plasmid Miniprep Kit (Sigma-Aldrich, USA). We then recombined these fragments into the Zebrafish Enhancer Detection ZED vector[65] following the protocol specified in[66]. Recombination into the ZED plasmid was performed taking into consideration the original orientation of the *fhl2b* genomic region. The resulting ZED plasmids were then purified with the DNA clean and concentrator -5 Kit (Zymo Research, USA). Injections were performed with 1 nl into one/two cell stage zebrafish (*Danio rerio*) embryos (*A. burtoni* construct was injected in wild-type strains AB and ABxEK, *P. philander* and *N. sexfasciatus* constructs were injected in wild-type strain ABxEK) with 25 ng/µl plasmid and 35 ng/µl Tol2 transposase mRNA. By outcrossing to wildtype zebrafish, we created five F2 stable transgenic lines for the *A. burtoni* construct, two F1 stable transgenic lines for the *P. philander* construct, and finally one F1 stable transgenic line for the *N. sexfasciatus* construct. Fish were raised and kept according to standard procedures[67]. Zebrafish were imaged using a Leica point scanning confocal microscope SP5-II-matrix and Zeiss LSM5 Pascal confocal microscope.

Fixation and Dehydration of Cichlid Fins

In order to determine the pigment cell composition of egg-spots (and especially if they contain iridophores in addition to xanthophores) we dissected *A. burtoni* anal fins. To better understand the morphological differences between non-haplochromine and haplochromine fins we further dissected three *N. crassus* anal fins. To visualize iridophores we removed the pteridine pigments of the overlying xanthophores by fixating the fin in 4%PFA-PBS for one hour at room temperature and washing it in a series of methanol:PBS dilutions (25%, 50%, 75%, 100%). Pictures were taken after six days in 100% methanol at -20 °C.

References

1. Pigliucci, M. What, if anything, is an evolutionary novelty? *Philos. Sci.* **75,** 887–898 (2008).
2. Carroll, S. B. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134,** 25–36 (2008).
3. Wagner, G. P. & Lynch, V. J. Evolutionary novelties. *Curr. Biol.* **20,** R48–52 (2010).
4. Wagner, A. The molecular origins of evolutionary innovations. *Trends Genet.* **27,** 397–410 (2011).
5. Carroll, S. B., Grenier, J. K. & Weatherbee, S. D. *From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design*. (Blackwell Science, 2001).
6. Schluter, D. Evidence for ecological speciation and its alternative. *Science* **323,** 737–41 (2009).
7. Lynch, V. J., Leclerc, R. D., May, G. & Wagner, G. P. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nat. Genet.* **43,** 1154–1159 (2011).
8. Beldade, P. & Brakefield, P. M. The genetics and evo-devo of butterfly wing patterns. *Nat. Rev. Genet.* **3,** 442–452 (2002).
9. W., Mack, T., Verheyen, E. & Meyer, A. Out of Tanganyika: genesis, explosive speciation, key-innovations and phylogeography of the haplochromine cichlid fishes. *BMC Evol. Biol.* **5,** 17 (2005).
10. Kocher, T. D. Adaptive evolution and explosive speciation: the cichlid fish model. *Nat. Rev. Genet.* **5,** 288–298 (2004).
11. Salzburger, W. The interaction of sexually and naturally selected traits in the adaptive radiations of cichlid fishes. *Mol. Ecol.* **18,** 169–185 (2009).
12. Fryer, G. & Iles, T. *The Cichlid Fishes of the Great Lakes of Africa: Their Biology and Evolution* (Oliver & Boyd, 1972).
13. Wickler, W. "Egg-dummies" as natural releasers in mouth-breeding cichlids. *Nature* **194,** 1092–1093 (1962).
14. Santos, M. E. & Salzburger, W. How Cichlids Diversify. *Science* **338,** 619–621 (2012).
15. Muschick, M., Indermaur, A. & Salzburger, W. Convergent evolution within an adaptive radiation of cichlid fishes. *Curr. Biol.* **22,** 2362–8 (2012).
16. Salzburger, W., Braasch, I. & Meyer, A. Adaptive sequence evolution in a color gene involved in the formation of the characteristic egg-dummies of male haplochromine cichlid fishes. *BMC Biol.* **5,** 51 (2007).
17. Braasch, I. & Postlethwait, J. in *Polyploidy Genome Evol.* (Soltis, D. E. & Soltis, P. S.) 341–383 (Springer, 2012).
18. Müller, J. M. *et al.* FHL2, a novel tissue-specific coactivator of the androgen receptor. *EMBO J.* **19,** 359–369 (2000).
19. Brun, J. *et al.* The LIM-only protein FHL2 controls osteoblast mesenchymal cell differentiation through non-canonical Wnt signalling. *Bone* **50,** S76 (2012).
20. Johannessen, M., Møller, S., Hansen, T., Moens, U. & Van Ghelue, M. The multifunctional roles of the four-and-a-half-LIM only protein FHL2. *Cell. Mol. Life Sci.* **63,** 268–284 (2006).
21. Kadrmas, J. L. & Beckerle, M. C. The LIM domain: from the cytoskeleton to the nucleus. *Nat. Rev. Mol. Cell Biol.* **5,** 920–931 (2004).
22. Heule, C. & Salzburger, W. The ontogenetic development of egg-spots in the haplochromine cichlid fish Astatotilapia burtoni. *J. Fish Biol.* **78,** 1588–1593 (2011).
23. Baldo, L., Santos, M. E. & Salzburger, W. Comparative transcriptomics of Eastern African cichlid fishes shows signs of positive selection and a large contribution of untranslated regions to genetic diversity. *Genome Biol. Evol.* **3,** 443–455 (2011).
24. Lynch, M. & Conery, J. S. The Evolutionary Fate and Consequences of Duplicate Genes. *Science* **290,** 1151–1155 (2000).

25. Takahashi, K., Terai, Y., Nishida, M. & Okada, N. A novel family of short interspersed repetitive elements (SINEs) from cichlids: the patterns of insertion of SINEs at orthologous loci support the proposed monophyly of four major groups of cichlid fishes in Lake Tanganyika. *Mol. Biol. Evol.* **15,** 391–407 (1998).
26. Britten, R. J. & Davidson, E. H. Gene regulation for higher cells: a theory. *Science* **165,** 349–57 (1969).
27. Britten RJ, D. E. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *Q. Rev. Biol.* **46,** 111–38 (1971).
28. Feschotte, C. Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.* **9,** 397–405 (2008).
29. Frohnhöfer, H. G., Krauss, J., Maischein, H.-M. & Nüsslein-Volhard, C. Iridophores and their interactions with other chromatophores are required for stripe formation in zebrafish. *Development* **140,** 2997–3007 (2013).
30. Patterson, L. B. & Parichy, D. M. Interactions with iridophores and the tissue environment required for patterning melanophores and xanthophores during zebrafish adult pigment stripe formation. *PLoS Genetics* **9,** e1003561 (2013).
31. Manousaki, T. *et al.* Parsing parallel evolution: ecological divergence and differential gene expression in the adaptive radiations of thick-lipped Midas cichlid fishes from Nicaragua. *Mol. Ecol.* **22,** 650–69 (2013).
32. Shimeld, S. M. & Holland, P. W. Vertebrate innovations. *Proc. Natl. Acad. Sci.USA* **97,** 4449–4452 (2000).
33. Egger, B., Klaefiger, Y., Theis, A. & Salzburger, W. A sensory bias has triggered the evolution of egg-spots in cichlid fishes. *PLoS One* **6,** e25601 (2011).
34. Theis, A., Salzburger, W. & Egger, B. The function of anal fin egg-spots in the cichlid fish Astatotilapia burtoni. *PLoS One* **7,** e29878 (2012).
35. Miller, S., Dykes, D. & Polesky, H. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucl. Acids Res.* **16,** 1215 (1988).
36. Meyer, B. S. & Salzburger, W. A novel primer set for multilocus phylogenetic inference in East African cichlid fishes. *Mol. Ecol. Resour.* **12,** 1097–104 (2012).
37. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437,** 376–80 (2005).
38. Binladen, J. *et al.* The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS One* **2,** e197 (2007).
39. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 3–5 (2011). doi:10.1093/bioinformatics/btq281.2
40. Kearse, M. *et al.* Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28,** 1647–9 (2012).
41. Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33,** 511–8 (2005).
42. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods* **9,** 772 (2012).
43. Schwarz, G. Estimating the Dimension of a Model. *Ann. Stat.* **6,** 461–464 (1978).
44. Ronquist, F. *et al.* MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61,** 539–542 (2012).
45. Zwickl, D. J. *Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion*. (University of Texas, Austin, 2006).
46. Sukumaran, J. & Holder, M. T. DendroPy: a Python library for phylogenetic computing. *Bioinformatics* **26,** 1569–71 (2010).
47. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

48. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).

49. Altschul, F. S., Gish, W., Miller, W., Myers, W. E. & Lipman, J. D. Basic Local Alignment Search Tool. *J. Mol. Biol.* **215,** 403–410 (1990).

50. Pfaffl, M. W. A new mathematical model for relative quantification in real-time RT-PCR. *Nucl. Acids Res.* **29,** e45 (2001).

51. Curran, K., Lister, J. A., Kunkel, G. R. Prendergast, A., Parichy, D. M. & Raible, D. W. Interplay between Foxd3 and Mitf regulates cell fate plasticity in the zebrafih neural crest. *Dev. Biol.* **344,** 107-118 (2010).

52. Tang, R., Dodd, A., Lai, D., McNabb, W. C & Love, D. R. Validation of zebrafish (*Danio rerio*) reference genes for quantitative real-time RT-PCR normalization. *Acta. Biochim. Biophys. Sin*. **39,** 384-390 (2007).

53. Zhang, Z. & Hu, J. Development and validation of endogenous reference genes for expression profiling of medaka (*Oryzias latipes*) exposed to endocrine disrupting chemicals by quantitative real-time RT-PCR. *Toxicol. Sci.* **95,** 356-368 (2007).

54. Catchen, J. M., Conery, J. S. & Postlethwait, J. H. Automated identification of conserved synteny after whole-genome duplication. *Genome Res.* **19,** 1497–505 (2009).

55. Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.* **132,** 365–386 (2000).

56. Flicek, P. *et al.* Ensembl 2012. *Nucl. Acids Res.* **40,** D84–90 (2012).

57. Larkin, M. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23,** 2947–2948 (2007).

58. Swofford, D. L. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). (Sinauer, Sunderland, 2003).

59. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24,** 1586–91 (2007).

60. Nielsen, R. & Yang, Z. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148,** 929–36 (1998).

61. Brawand, D. *et al.* The genomic substrate for adaptive radiation in Africa cichlid fish. *Nature* (in press)

62. Pages, H., Aboyoun, P., Gentleman, R. & DebRoy, S. Biostrings: String objects representing biological sequences, and matching algorithms. R-package

63. Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M. & Dubchak, I. VISTA: computational tools for comparative genomics. *Nucl. Acids Res.* **32,** W273–279 (2004).

64. Brudno, M. *et al.* LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* **13,** 721–731 (2003).

65. Bessa, J. *et al.* Zebrafish enhancer detection (ZED) vector: a new tool to facilitate transgenesis and the functional analysis of cis-regulatory regions in zebrafish. *Dev. Dyn.* **238,** 2409–2417 (2009).

66. Bessa, J. & Gómez-Skarmeta, J. L. Making Reporter Gene Constructs to Analyze Cis - regulatory Elements. *Mol. Methods Evol. Genet.* **772,** 397–408 (2011).

67. Westerfield, M. *The zebrafish book. A guide for the laboratory use of zebrafish (Danio rerio)* (University of Oregon Press, 2000).

Acknowledgments

We thank A. Theis, T. Bosia, B. Egger, G. Banyankimbona, E. Schraml, M. Muschick and H. Gante for assistance with the photographs; P. Beldade, J. Ganz, H. Hoekstra, J. H. Postlethwait, G. Wagner and the members of the Salzburger Lab for discussions; J. Bessa for sending us the ZED vector; V. Laudet for providing medaka and zebrafish; J. Ganz, University of Oregon, and the Imaging Core Facility of the Biozentrum, University of Basel for support with confocal microscopy; the University of Oregon zebrafish facility, A. Indermaur and A. Rüegg (Basel) for fish keeping; and the BROAD institute and the Cichlid Genome Consortium for sharing unpublished data on the cichlid genomes. This study was supported by grants from the Fundação para a Ciência e Tecnologia (FCT, Portugal) to M.E.S.; the Volkswagen Stiftung to I. B. and A. B.; the "Fellowship for Excellence" from the International PhD Program in Molecular Life Sciences and the Werner-Siemens-Foundation (Aug) to L.S.; the Swiss National Science Foundation (SNF) to M. A.; and the European Research Council (ERC, Starting Grant 'INTERGENADAPT'), the University of Basel, the National Geographic Society, and the Swiss National Science Foundation (SNF) to W. S. The work of I. B. was further supported by NIH grant R01 RR020833 (R01 OD011116) to John H. Postlethwait.

Author's Contributions:

M.E.S., I.B. and W.S. designed the study. M.E.S and W.S. collected the samples, M.E.S. performed the RNAseq, gene expression, comparative genomics and zebrafish functional analysis. N.B. performed the sequencing of *fhl2* paralogs coding region and analyzed its rates of evolution. B.S.M. collected the 454 sequence data and B.S.M. and W.S. performed the phylogenetic analysis. A.B. performed the SINE consensus alignments and analyzed the transposable element composition of *fhl2b* genomic region. I.B. performed the zebrafish functional assays of the *A. burtoni* construct and *fhl2* paralogs synteny analysis. L.S., H-G.B. and M.A. assisted with the zebrafish functional assays of the *A.burtoni*, *P. philander* and *N. sexfasciatus* construct. M.E.S. and W.S. wrote the paper and all authors contributed to revisions.

Additional information

Accession codes: All nucleotide sequences reported in this study were deposited at GenBank (accession numbers KM263618 to KM264016). All the short reads were deposited at the Sequence Read Archive (SRA) (BioProject ID PRJNA25755).

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Chapter 4

Transcriptomics of a novel and variable pigment trait in cichlid fishes – identification of candidate genes for egg-spot morphogenesis

M. Emília Santos and Walter Salzburger

Transcriptomics of a novel and variable pigment trait in cichlid fishes – identification of candidate genes for egg-spot morphogenesis

M. Emília Santos and Walter Salzburger

Zoological Institute, University of Basel, Vesalgasse 1, 4051 Basel, Switzerland,

Correspondence: Walter Salzburger, Fax: +41 61 267 0301: E-mail: walter.salzburger@unibas.ch

Keywords: evolutionary novelty, diversification, egg-spot, East African Cichlids, RNAseq, qPCR

4.1 – Abstract

Background: Understanding the genetic basis of novel traits is a hot topic in evolutionary biology. Egg-spots are circular pigmentation markings on the anal fins of haplochromine cichlid fishes. They are a novel and variable trait and play an important role in the breeding behavior of this group of fishes. Our knowledge about the underlying genetics of this trait is sparse. With next-generation sequencing it is now possible to access transcriptomes and gene expression at unprecedented levels in order to generate candidate genes. Here we present an RNAseq survey of *Astatotilapia burtoni* egg-spot and anal fin transcriptome.

Results: We present a differential gene expression analysis between adult egg-spot and anal fins transcriptomes. We generated candidate genes and identified them using BLAST and characterized their functions using a Gene Ontology database. We further confirmed some of these candidates as egg-spot genes via inter-species gene expression comparisons using qPCR. Among these egg-spot genes are previously known patterning genes, such as *hoxC12a* and *bmp3,* and pigmentation related genes, eg. *asip1*. Surprisingly, we detected many (~30%) egg-spot contigs that we were unable to identify via similarity searches suggesting that these are linage (cichlid) specific genes.

Conclusions: We provide evidence that both co-option of pre-existing genes and lineage specific genes might be involved in the formation of the egg-spot phenotype and suggesting that both mechanisms play a role in the evolution of novel traits. Finally, we have identified a set of candidate genes that will serve as an important and useful resource for future research on the emergence and diversification the egg-spot trait.

4.2 - Background

How novel traits, i.e. lineage-specific traits that perform new functions, emerge and are modified is one of the many unresolved problems in evolutionary biology [1–3]. The bulk of evidence so far suggests that morphological novelties emerge largely through co-option of pre-existing regulatory networks via changes in gene *cis*-regulation ("old genes play new tricks")[4, 5]. However, recent work suggests that structural mutations, gene duplication and lineage specific genes might also play a role [6–9]. The genetics underlying phenotypic innovation and variation is still an open question and intensely debated [10–14]. The main problem with addressing this question, or discern between the alternative hypotheses, is that most research has been done with single gene case studies, comparing species that are too far apart phylogenetically, or that are in lineages that do not show much diversity in the trait in question. By comparing the gene expression behaviour of several genes in different species with different versions of the same phenotype, one might be able to address such types of questions.

Animal colour patterns are a highly intra- and inter-specific variable phenotype, making pigmentation a suitable trait to study the genetics of morphological diversification. Furthermore, it is easy to assess colour patterns functionality because these traits evolve as an adaptation to the surrounding environment via natural selection (inter- and intra-specific communication, camouflage and mimicry), or co-vary with female choice via sexual selection [15]. The outcome of these two types of selection is often very different, with the first one generating cryptic phenotypes, where colouration mimics the environment, and the second case generating conspicuous phenotypes (badge of status), where generally males show off beautiful colours in order to be chosen by females. The study of pigment patterns has already demonstrated its power and potential contributions to the understanding of the genetics of the emergence of novel traits, diversification and adaptation (reviewed in [16–18]).

East African cichlid fishes are the most species rich extant vertebrate family and represent a major model for diversification and speciation [19–21]. Together Lake Tanganyika, Lake Malawi and Lake Victoria contain the largest set of cichlid species (~2000 species), which are thought to have evolved in the last few million to thousands of years only, i.e. in a very short period of time. These hundreds of species are thriving with different arrays of colour and pigment patterns. The three lakes represent different stages of species divergence and phenotypic differentiation, thus providing an ideal set up to study pigment pattern diversification and adaptation in an extremely large number of species and in three replicate divergence timescales [19–21].

The haplochromines represent the most species-rich group of cichlids (~1500 species). What distinguishes this lineage from other, less taxonomically diverse, cichlid lineages is maternal mouthbrooding, together with the conspicuous male pigment patterns, including anal fin egg-spots [22]. Egg-spots are circular markings in the anal fins of the haplochromine males, consisting of a central circular area of xanthophore surrounded by an outer transparent ring. They play a key role in the territorial and breeding behaviour of this group of maternal mouthbrooding fishes. Their function is variable, being sexually selected via female choice in some species [23, 24] and via male-male competition in others [25, 26]. This trait is also an important component of the courtship behaviour, as described in Salzburger *et al.* 2007 [27]. In some species, females also show egg-spots but they are usually less conspicuous than in males. Egg-spots emerged only once along with the origin of haplochromines and are thought to have influenced the evolutionary success of this cichlid lineage. Egg-spots show an extreme inter- and intra-specific variability in different numbers, colours and positions on the fin, and are therefore an ideal case study to address origin, diversification and ongoing adaptation of novel traits [28].

Pigmentation diversity in fish is determined by the arrangement, position, density, and specification of different pigment cell types [29]. These characteristics, in turn, depend on other factors such as neural crest cell migration, specification, proliferation, and survival. Work in other fish model systems has shed some light in the genes involved in these processes [30]. Several studies have addressed pigmentation diversity in East African cichlids but little is known about the developmental pathways underlying colouration and pigmentation patterning. So far, only a few genes have been studied. One of these genes is *hagoromo,* which shows a greater diversity of alternatively spliced variants and accelerated protein evolution in the haplochromines [31, 32], and the *paired box 7* which are linked to a haplochromine female biased pigmentation phenotype [33]. When considering the egg-spot phenotype itself, little advance has been made. The xanthophore marker colony stimulating factor 1 receptor A (*csf1ra*) underwent adaptive sequence evolution in the ancestral lineage of the haplochromines coinciding with the emergence of egg-dummies and has shown to be involved in the formation of the egg-spots [27]. However, *csf1ra* is rather downstream in the morphogenesis of egg-spots and many more upstream genes would need to be studied to understand the evolution and developmental basis of egg-spots in cichlids. Recently, we showed that the two *four and a half lim domain 2* proteins (*fhl2a* and *fhl2b*) are also involved in egg-spot development and that an alteration in the cis-regulatory region of *fhl2b* could potentially have contributed to the emergence of this trait in haplochromines (chapter 3). This is exciting progress but genes do not act alone; they interact with other genes and with the environment in order to build a phenotype. If we want to understand how traits emerge and diversify we have to understand not only one gene and its phenotypic effect, but also the pathway where it is integrated and its developmental role. Unfortunately, our knowledge on the genes that might underlie this phenotype is limited.

With the development of next generation sequencing techniques it is now possible to study genomes, transcriptomes and gene expression at unprecedented levels [34, 35]. With the aim of elucidating the mechanisms of origin and diversification of the egg-spot trait, we took a transcriptomic approach to generate candidate genes involved in the formation of the trait. We measured gene expression level in the egg-spot from the haplochromine *Astatotilapia burtoni*, using the non-pigmented region of the anal fin as a reference (figure 1A). We then characterized this transcriptome both quantitatively and qualitatively, generating functional relevant groups and a database of potential egg-spot candidate genes. We further examined the expression of the 24 most overexpressed and the 24 most underexpressed genes in another haplochromine species, *Pseudothropheus pulpican,* using qPCR. *P. pulpican* has egg-spots in a different position on the anal fin, and therefore this assay was used to confirm the involvement of these genes in the egg-spot, rather than involved in fin patterning. We repeated the same assay in *Callochromis macrops*, a species belonging to another lineage – the ectodines. This species has no egg-spots in its anal fin, instead possesses a “blotch” of pigmentation with non-discrete/ill-definied boundaries. Even though ectodine blotch and haplochromine egg-spot are not homologous, studying gene expression in both traits can help to shed light in what are the egg-spot specific genes and how similar the genetic basis of both traits is. With this strategy we describe a novel set of potential egg-spot gene candidates. The genes identified here will be studied further and hopefully will help shed light on the emergence and evolution of the haplochromine egg-spot, as well as novel traits in general.

4.3 - Results

Transcript profile in anal fin and egg-spot tissue

In order to identify genes involved in egg-spot morphogenesis we conducted an RNAseq experiment, which involved quantifying differences in gene expression between egg-spots and the rest of the non-pigmented anal fin of *Astatotilapia burtoni* (figure 1A). Anal fin tissue was used as a reference so we could measure both egg-spot transcript over-expression and under-expression in relation to anal fin. In total 193,054,988 high quality reads were obtained from egg-spot samples and 194,099,061 from anal fin samples from six individuals.

We mapped the reads from each tissue to a reference *A. burtoni* embryonic transcriptome that contains 171,136 reference transcripts. We chose this reference because it is a collection from several different embryonic and larval developmental stages, and therefore probably the most comprehensive available representation of the entire gene set from *A. burtoni*. This trancriptome was sequenced and provided by Broad Institute (Massachusetts, USA), and is part of the cichlid genome project (http://www.broadinstitute.org/models/tilapia). The reference transcriptome is a redundant database, meaning that it includes splice-variants and allelic variants for each gene. These transcripts are annotated and can be traced back to a common unique contig (see methods).

We compared the expression between the two tissues – egg-spot and non-pigmented anal fin – in six replicates. In total we found that 7428 transcripts were differentially expressed (figure 1B). The differentially expressed transcripts, together with the respective expression values, can be found in supplementary file 1. When we concatenate this dataset to its parent contigs (instead of transcripts) there is differential expression of 4259 contigs – 1139 overexpressed in the egg-spot and 3120 underexpressed (figure 1B).

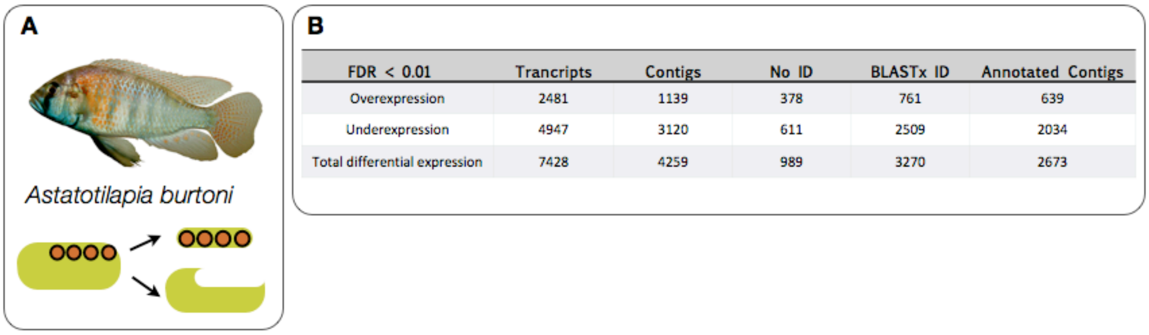
Functional annotation

We annotated the 7428 differentially expressed transcripts using blast2GO [36]. First, we performed a BLASTx search against NCBI's non-redundant database for each differentially expressed sequence, then we reduced and concatenated the transcript dataset annotations to the parent contig (both annotated datasets can be found in supplementary file 2). From the 4259 differential expressed contigs, 3270 had positive BLAST hits against the database. From these 3270 contigs we could functionally annotate 2673 (with blast2GO and interproscan; figure 1B). The 989 contigs without positive BLAST hits could represent non-coding RNAs, lineage specific genes (new or fast evolving genes), or partial sequences of known genes that could not be identified. Surprisingly, around 30% (378/1139) of the egg-spot overexpressed contigs were non-identified.

In figure 2 we describe the Gene Ontology (GO) term composition for two datasets: egg-spot overexpression (dataset A) and egg-spot underexpression (dataset B). Overall, the GO level 2 term representation was significantly different between the two datasets (figure 2), showing that there are functionally very different (Biological process: χ^2=179.3324715, *df* = 13, *p* < 0.0001; Molecular process: χ^2=68.45, *df* = 9, *p* < 0.0001; Cellular process: χ^2=68.66, *df* = 4, *p* < 0.0001).

With the aim to further functionally describe our two datasets, and to generate potential candidate genes of interest, we tested for differential enrichment of all the GO level terms represented in dataset A relative to dataset B (supplementary file 3).

Figure 1– **A)** Schematic representation of the RNAseq experimental design. Egg-spots were separated from the rest of the anal fin. The transcriptomes from both tissues were then sequenced with the remaining anal fin being used as reference. **B)** Differential gene expression statistics. Number of genes overexpressed and underexpressed in the egg-spot are shown.



FDR < 0.01	Trancripts	Contigs	No ID	BLASTx ID	Annotated Contigs
Overexpression	2481	1139	378	761	639
Underexpression	4947	3120	611	2509	2034
Total differential expression	7428	4259	989	3270	2673

Depending on the developmental or physiological perspective one wants to study, one can choose a differentially represented term (whether enriched or underrepresented) and characterize the genes belonging to this functional category (GO term). In table 1 we show four of 205 functional categories that were enriched in the egg-spot overexpressed dataset (dataset A). The GO category ‘pigmentation’ (GO:0043473) was chosen for obvious reasons, ‘neural crest cell differentiation’ (GO:0014033) and ‘cell motility’ (GO:0048870) were chosen because neural crest cells are precursors of pigment cells [37, 38].

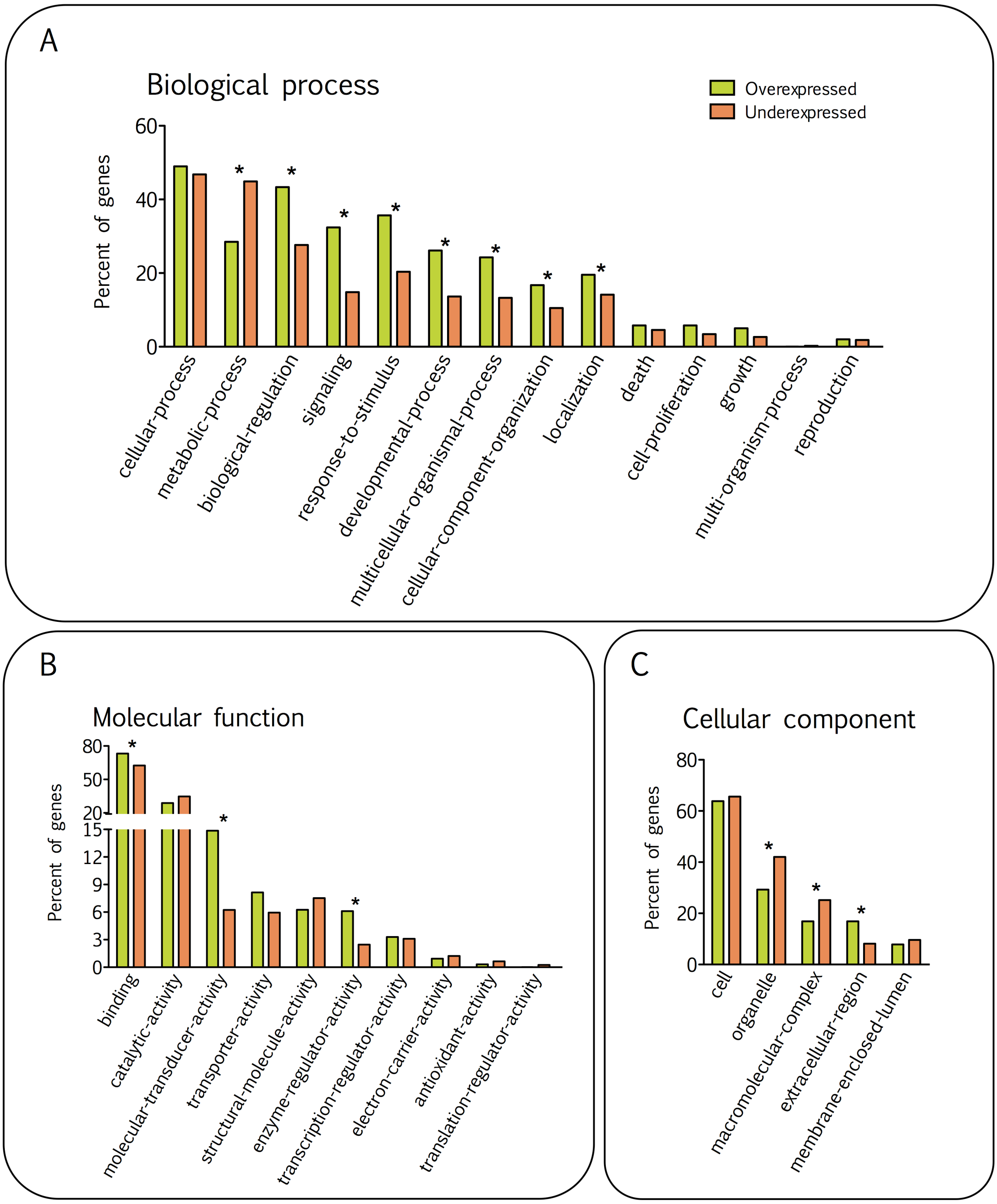


Figure 2– Gene ontology (GO) ID representations for both our datasets. Egg-spot overexpression is depicted in green and underexpression is depicted in red. **A)** Biological process ontology **B)** Molecular function ontology **C)** Cellular component ontology. Asterik marks denotes significant difference in proportion of genes between the two datasets as shown by chi-squared test (corrected for multiple tests).

Table 1 – Egg-spot enriched GO terms and potential candidate genes to be further studied

GO term	Description	Egg-spot candidate genes	
GO:0043473	**Pigmentation**	mitfa	microphthalmia-associated transcription factor a
		ednrb1	endothelin receptor b1
		pax7	paired box 7
		alk	alk tyrosine kinase receptor
		dock7	dedicator of cytokinesis 7
		sox10	sex determining region Y box 10
		myo5a	myosin 5a
		csf1ra	colony stimulating factor 1 receptor a
		pmela	silva
		alk	alk tyrosine kinase receptor
GO:0014033	**Neural crest cell differentiation**	irx1	iroquois homeobox 1
		ednrb1	endothelin receptor b1
		pax7	paired box 7
		tnc	tenascin c
		hand2	heart and neural crest derivatives-expressed 2
		noe2	noelin 2
		sema3fa	semaphorin 3 fa
		rdh10	retinol dehydrogenase 10
		sox10	sex determining region Y box 10
GO:0007169	**Transmembrane receptor protein tyrosine kinase signalling pathway**	msp	hepatocyte growth factor-like
		ddr2	discoidin domain-containing receptor 2
		pkca	protein kinase c alpha
		pik3r1	phosphatidylinositol 3-kinase regulatory subunit alpha
		pax7	paired box 7
		alk	alk tyrosine kinase receptor
		sorbs	novel protein vertebrate sorbin and sh3 domain containing family
		hel	replicase/helicase/endonuclease
		alk	alk tyrosine kinase receptor
		spry2	sprouty 2
		nrcam	neuronal cell adhesion molecule
		csf1r	macrophage colony-stimulating factor receptor
		enpp1	ectonucleotide pyrophosphatase phosphodiesterase family member 1
		IgSF10	immunoglobulin superfamily member 10
		csf1ra	macrophage colony-stimulating factor receptor 1a
		vgefr2b	kinase insert domain receptor b
		flt4	fms-like tyrosine kinase 4
GO:0016477	**Cell migration**	tbr2	eomesodermin
		limx1b	lim homeobox transcription factor beta 1
		plxna2	plexin a2
		pkca	protein kinase c alpha
		muc2	mucin-2
		sdf1	stromal cell-derived factor 1
		ednrb1	endothelin receptor b1
		vangl2	vang-like protein 2
		lama4	laminin subunit alpha 4
		itga1	integrin alpha 1
		dab1	disabled homolog 1
		pik3r1	phosphatidylinositol 3-kinase regulatory subunit alpha
		tnc	tenascin c
		myh	nonmuscle myosin heavy chain
		tgfbr3	transforming growth factor beta receptor 3
		ncam	neuronal cell adhesion molecule
		nlrp12	lrr and pyd domains-containing protein 12
		chrd	chordin
		nr1	nuclear receptor subfamily group member 1
		csf1r	macrophage colony-stimulating factor receptor
		prkg1	cgmp-dependent protein kinase 1-like
		sema3fa	semaphorin 3 fa
		ntn1	netrin1
		plxna2	plexin a2
		astn1	astrotactin 1
		cool1	rho guanine nucleotide exchange factor 7
		gpc3	glypican 3
		nr2f2	coup transcription factor 2
		sox10	sex determining region Y box 10
		cxcr7	c-x-c chemokine receptor 7
		itga1	integrin alpha

Neural crest cells have to migrate from their original location to the anal fin where they will form the egg-spots [39]. Egg-spots will only be formed though if there is pigment production, which in turn is often activated via tyrosine kinases [40–42]. Therefore, we also selected 'transmembrane receptor protein tyrosine kinase signaling pathway' (GO:0007169) as an informative functional categories to extract candidate genes. One gene that was highlighted using this method, *csf1ra*, was shown previously to be involved in egg-spot development [27]. This result demonstrates that our strategy is a good approach to generate candidate genes. This is a supervised search however, and will have biases, because one chooses a candidate for what is already known. There are many other non-described genes, or known genes with incomplete GO term annotations that could play a role in egg-spot morphogenesis. Therefore, we characterized the expression of some of the most differentially expressed genes further by means of qPCR.

Egg-spot associated transcripts and candidate genes

We further examined the 24 most differentially expressed genes from both the egg-spot overexpressed (dataset A) and underexpressed lists (dataset B) by means of qPCR (see table 2 and table 3 for a list of genes overexpressed and underexpressed genes respectively). Among them there are five overexpressed contigs and two underexpressed contigs that remained unidentified after BLASTx and BLASTn searches against a non-redundant NCBI database. Overall, there was no obvious trend in the functional categories associated with these top differentially expressed genes, ranging from signaling, structural proteins, to transcription factors (data not shown, available in supplementary table 2). These genes are differentially expressed between the egg-spot tissue and non-pigmented anal fin tissue from *Astatotilapia burtoni*. This expression may be egg-spot correlated, but can as well be correlated just with the distal region of the fin.

Gene expression in *Pseudotropheus pulpican*

Egg-spots of *A. burtoni* are located in the proximal region of the fin (figure 3A). Therefore, expression in this region can be related with the proximal region of the anal fin and not with egg-spots. To further validate these genes as participating in the egg-spot morphogenesis, we studied their expression in another species *Pseudotropheus pulpican* (figure 3A). The haplochromine *P. pulpican* has its egg-spots in a different position in the anal fin, and therefore we can control for fin patterning gene expression.

Table 2 – Top 24 differentially **overexpressed** transcripts and their identification as determined through BLASTx agains NCBI non-redundant database. Gene order is defined difference in expression.

Egg-spot overexpression					BLASTx Identification		
	Gene	Transcript	logFC	p-value	Description	Accession	e-value
1	asip1	comp13033_c0_seq1	3.1014521929	7.19E-113	PREDICTED: agouti-signaling protein-like [Oreochromis niloticus]	XP_003448419.1	3.00E-25
2	*NA*	comp17910_c0_seq1	2.4077958344	1.64E-74	No significant similarity found	*NA*	*NA*
3	rbp7	comp8091_c0_seq1	3.1134760208	1.03E-67	PREDICTED: retinoid-binding protein 7-like [Oreochromis niloticus]	XP_003448369.1	9.00E-91
4	IF ON3	comp1238_c0_seq1	2.1918579644	1.94E-67	PREDICTED: intermediate filament protein ON3-like [Oreochromis niloticus]	XP_003441441.1	0
5	akap12	comp28860_c0_seq1	2.2994280142	5.54E-55	A-kinase anchor protein 12 [Danio rerio] >gb|ABQ11279.1| gravin [Danio rerio]	NP_001091654.1	2.00E-49
6	*NA*	comp20229_c0_seq1	2.6007126777	4.27E-52	PREDICTED: hypothetical protein LOC100708826 [Oreochromis niloticus]	XP_003455230.1	6.00E-19
7	bmp3b	comp14170_c0_seq1	1.8473666295	6.40E-47	PREDICTED: bone morphogenetic protein 3B-like [Oreochromis niloticus]	XP_003438593.1	0
8	rbp4a	comp104_c0_seq1	1.697872898	2.57E-44	PREDICTED: retinol-binding protein 4-A-like [Oreochromis niloticus]	XP_003441907.1	2.00E-132
9	cytl1	comp7733_c0_seq1	1.6965569195	3.05E-40	PREDICTED: cytokine-like protein 1-like [Oreochromis niloticus]	XP_003441598.1	4.00E-80
10	*NA*	comp23699_c0_seq1	2.0393467074	7.88E-38	No significant similarity found	*NA*	*NA*
11	*NA*	comp4443_c1_seq1	1.6042281652	5.90E-36	No significant similarity found	*NA*	*NA*
12	fhl2a	comp2939_c0_seq1	1.480050043	2.51E-35	PREDICTED: four and a half LIM domains protein 2-like [Oreochromis niloticus]	XP_003453001.1	0
13	hand2	comp22787_c0_seq1	2.5349400584	3.54E-35	PREDICTED: heart- and neural crest derivatives-expressed protein 2-like [Oreochromis niloticus]	XP_003452793.1	2.00E-96
14	*NA*	comp23328_c0_seq1	2.4425132894	1.58E-34	No significant similarity found	*NA*	*NA*
15	cecr5	comp6479_c0_seq1	1.4571347609	2.43E-33	PREDICTED: cat eye syndrome critical region protein 5-like [Oreochromis niloticus]	XP_003457763.1	0
16	mucin2	comp3522_c2_seq4	1.4745684131	1.85E-32	PREDICTED: mucin-2-like [Danio rerio]	XP_003201446.1	4.00E-07
17	sfr5	comp6979_c0_seq1	1.5784370516	2.48E-31	PREDICTED: secreted frizzled-related protein 5-like isoform 3 [Oreochromis niloticus]	XP_003451970.1	0
18	igf1	comp17864_c0_seq1	1.4333096407	6.91E-30	PREDICTED: insulin-like growth factor 1 [Oreochromis niloticus]	XP_003448107.1	7.00E-94
19	zygin1	comp2115_c0_seq1	1.4864441669	7.31E-30	PREDICTED: fasciculation and elongation protein zeta-1-like [Oreochromis niloticus]	XP_003449843.1	0
20	vtn	comp7947_c0_seq1	1.421380662	7.48E-30	PREDICTED: vitronectin-like [Oreochromis niloticus]	XP_003458657.1	0.00E+00
21	*NA*	comp24816_c0_seq1	1.6201728647	3.78E-27	No significant similarity found	*NA*	*NA*
22	hoxC12a	comp21426_c0_seq1	1.8657519852	1.81E-26	Hoxc12a [Haplochromis burtoni]	ABS70754.1	2.00E-172
23	igSF10	comp36206_c0_seq1	1.3856487322	1.77E-25	PREDICTED: immunoglobulin superfamily member 10-like [Oreochromis niloticus]	XP_003454869.1	0
24	fmdo	comp19154_c0_seq2	1.2845536844	3.94E-25	PREDICTED: fibromodulin-like [Oreochromis niloticus]	XP_003441412.1	0

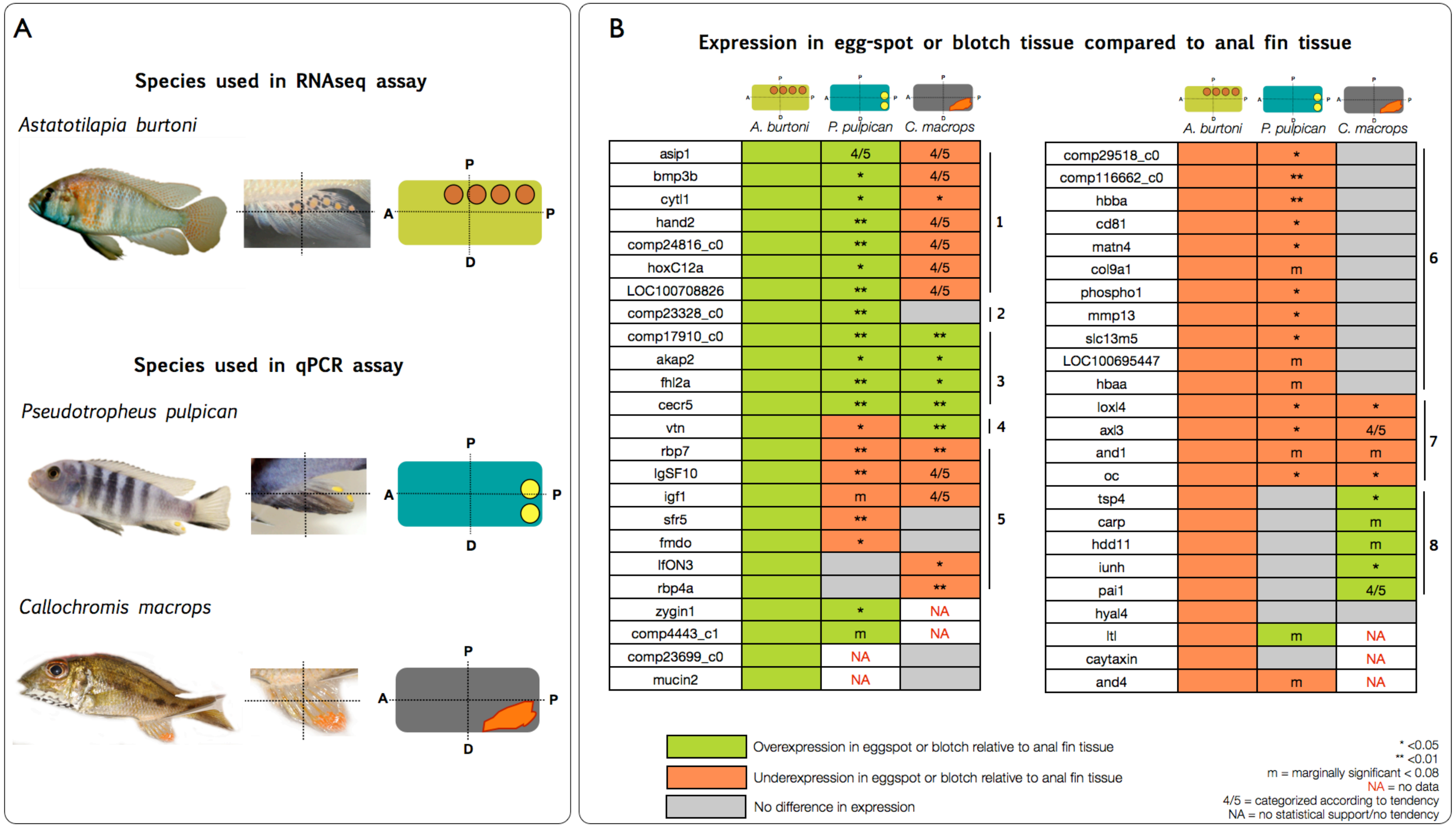
Table 3 – Top 24 differentially **underexpressed** transcripts and their identification as determined through BLASTx agains NCBI non-redundant database. Gene order is defined by difference in expression

	Egg-spot underexpression				BLASTx Identification		
	Gene	**Transcript**	**logFC**	**p-value**	**Description**	**Accession**	**e-value**
1	axl3	comp20108_c0_seq1	-5.023778111	2.15E-154	PREDICTED: homeobox protein aristaless-like 3-like [Danio rerio]	XP_695330.1	2.00E-152
2	and1	comp5622_c0_seq1	-3.076677029	6.29E-136	actinodin1 precursor [Danio rerio]	NP_001184183.1	4.00E-124
3	oc	comp5530_c0_seq1	-3.075997615	5.99E-131	PREDICTED: osteocalcin [Oreochromis niloticus]	XP_003443144.1	2.00E-62
4	slc13m5	comp28513_c0_seq1	-3.194127439	1.26E-116	solute carrier family 13, member 5 [Danio rerio]	NP_001136038.1	0
5	and4	comp2301_c0_seq1	-2.757621569	3.80E-101	actinodin4 precursor [Danio rerio]	NP_001129716.1	1.00E-85
6	carp	comp10574_c0_seq1	-2.81183119	1.62E-84	PREDICTED: cocaine- and amphetamine-regulated transcript protein-like [Oreochromis niloticus]	XP_003456941.1	3.00E-58
7	*NA*	comp36289_c0_seq1	-3.46607511	1.11E-72	PREDICTED: hypothetical protein LOC100695447 [Oreochromis niloticus]	XP_003459280.1	2.00E-50
8	hdd11	comp1748_c0_seq1	-2.211148193	1.48E-65	PREDICTED: putative defense protein Hdd11-like [Oreochromis niloticus]	XP_003446154.1	8.00E-127
9	*NA*	comp116662_c0_seq1	-2.670341607	2.85E-60	No significant similarity found	*NA*	*NA*
10	*NA*	comp29518_c0_seq1	-2.523239461	4.03E-57	No significant similarity found	*NA*	*NA*
11	hbba	comp70_c0_seq1	-1.802167206	6.22E-50	PREDICTED: hemoglobin subunit beta-A-like isoform 1 [Oreochromis niloticus]	XP_003442119.1	9.00E-99
12	iunh	comp29726_c0_seq1	-2.029292678	5.83E-48	PREDICTED: inosine-uridine preferring nucleoside hydrolase-like [Oreochromis niloticus]	XP_003455949.1	6.00E-55
13	tsp4	comp2186_c1_seq1	-1.701105028	1.59E-47	PREDICTED: thrombospondin-4-B-like [Oreochromis niloticus]	XP_003451568.1	0
14	col9a1	comp6219_c0_seq1	-1.666407343	1.01E-45	PREDICTED: collagen alpha-1(IX) chain-like, partial [Danio rerio]	XP_003200573.1	2.00E-138
15	phospho1	comp2411_c0_seq1	-1.60077492	3.46E-42	PREDICTED: probable phosphatase phospho1-like [Oreochromis niloticus]	XP_003442063.1	0
16	mmp13	comp20376_c0_seq1	-1.990458013	1.35E-41	PREDICTED: collagenase 3-like [Oreochromis niloticus]	XP_003441718.1	0
17	ltl	comp656_c0_seq1	-1.54486822	4.06E-41	lily-type lectin [Epinephelus coioides]	AEA39736.1	3.00E-69
18	pai1	comp29400_c0_seq1	-1.70804808	5.92E-41	PREDICTED: plasminogen activator inhibitor 1-like [Oreochromis niloticus]	XP_003460165.1	0
19	loxl4	comp12727_c0_seq1	-1.527019208	1.10E-39	PREDICTED: lysyl oxidase homolog 4-like [Oreochromis niloticus]	XP_003455871.1	0.00E+00
20	cd81	comp5209_c0_seq1	-1.550977998	2.92E-39	PREDICTED: CD81 antigen-like [Oreochromis niloticus]	XP_003443898.1	0.00E+00
21	matn4	comp4244_c0_seq1	-1.730976159	3.62E-39	PREDICTED: matrilin-4 [Oreochromis niloticus]	XP_003451941.1	0
22	hbaa	comp28_c0_seq1	-1.56634155	1.41E-38	RecName: Full=Hemoglobin subunit alpha-A	Q9PVM4.3	1.00E-79
23	caytaxin	comp7321_c0_seq1	-1.662999354	9.78E-38	PREDICTED: caytaxin-like [Oreochromis niloticus]	XP_003448582.1	0
24	hyal4	comp645_c1_seq5	-2.243665185	3.58E-36	PREDICTED: hyaluronidase-4-like [Oreochromis niloticus]	XP_003449274.1	0

If an overexpressed gene in *A. burtoni* is related to fin patterning only, then this same gene should be underexpressed in the egg-spot of *P.pulpicans* and *vice versa*. We decided to further study egg-spot underexpression, and not just overexpression, because there could be an egg-spot inhibitor that does not allow the development of pigmentation in other regions of the anal fin. The expression of the unidentified contigs was also examined in order to dissect their function. Egg-spots in *A. burtoni* and *P. pulpican* can be considered homologous as the trait evolved once in this lineage [22]. The fact that they are homologous does not necessarily mean that the trait shares exactly the same genetic basis in each species, and there will be lineage specific egg-spot genes that will be missed with this strategy.

Figure 3B (left panel) demonstrates the results for egg-spot overexpression dataset. For this dataset 14 out of 24 genes showed overexpression in the egg-spots of *P. pulpican*. Interestingly, the non-identified contigs were also overexpressed in the egg-spots suggesting that they are functionally relevant and probably play a role in the morphogenesis of egg-spot. Among the genes that are overexpressed in both types of egg-spots, there are two well known transcription factors known to be involved in patterning and cell fate specification (*hoxC12a, hand2* respectively*)*, and a well known growth morphogen (*bmp3b*) [43–45]. Another interesting candidate is *asip1*, which in mammals is connected with the development of lighter skin coats [46]. Eight out of the 24 genes are underexpressed, or show no difference in expression in the egg-spots of *P. pulpican*, suggesting that these genes are not involved in egg-spot morphogenesis but are involved in fin patterning. We do not rule out that these genes might also be responsible for interspecific differences of the egg-spot phenotype, acting in a lineage specific manner.

In figure 3B (right panel) are the results of the expression study for the underexpressed data set in *A. burtoni*. The rationale behind the experiment was the same. If the underexpression of a gene is correlated with the presence of the egg-spots than it should be underexpressed in the egg-spots of *P. pulpican,* even though they are in a different position in the fin. For this dataset 16 out of 24 genes showed underexpression in egg-spots of *P. pulpican,* including the non-identified contigs, meaning that underexpression is egg-spot related. In both qPCR datasets (A and B) there were genes for which we did not have enough statistical power (or an obvious trend in expression) to decide whether they would be over- or underexpressed, this is an ongoing project and five more individual replicates will be tested for this experiment.


A
Species used in RNAseq assay
Astatotilapia burtoni
P
A
P
D
Species used in qPCR assay
Pseudotropheus pulpican
P
A
P
D
Callochromis macrops
P
A
P
D
B
Expression in egg-spot or blotch tissue compared to anal fin tissue

	A. burtoni	P. pulpican	C. macrops	
asip1	🟩	🟩 4/5	🟧 4/5	1
bmp3b	🟩	🟩 *	🟧 4/5	
cytl1	🟩	🟩 *	🟧 *	
hand2	🟩	🟩 **	🟧 4/5	
comp24816_c0	🟩	🟩 **	🟧 4/5	
hoxC12a	🟩	🟩 *	🟧 4/5	
LOC100708826	🟩	🟩 **	🟧 4/5	
comp23328_c0	🟩	🟩 **	⬜	2
comp17910_c0	🟩	🟩 **	🟩 **	3
akap2	🟩	🟩 *	🟩 *	
fhl2a	🟩	🟩 **	🟩 *	
cecr5	🟩	🟩 **	🟩 **	
vtn	🟩	🟧 *	🟩 **	4
rbp7	🟩	🟧 **	🟧 **	5
IgSF10	🟩	🟧 **	🟧 4/5	
igf1	🟩	🟧 m	🟧 4/5	
sfr5	🟩	🟧 **	⬜	
fmdo	🟩	🟧 *	⬜	
IfON3	🟩	⬜	🟧 *	
rbp4a	🟩	⬜	🟧 **	
zygin1	🟩	🟩 *	NA	
comp4443_c1	🟩	🟩 m	NA	
comp23699_c0	🟩	NA	⬜	
mucin2	🟩	NA	⬜	

	A. burtoni	P. pulpican	C. macrops	
comp29518_c0	🟧	🟧 *	⬜	6
comp116662_c0	🟧	🟧 **	⬜	
hbba	🟧	🟧 **	⬜	
cd81	🟧	🟧 *	⬜	
matn4	🟧	🟧 *	⬜	
col9a1	🟧	🟧 m	⬜	
phospho1	🟧	🟧 *	⬜	
mmp13	🟧	🟧 *	⬜	
slc13m5	🟧	🟧 *	⬜	
LOC100695447	🟧	🟧 m	⬜	
hbaa	🟧	🟧 m	⬜	
loxl4	🟧	🟧 *	🟧 *	7
axl3	🟧	🟧 *	🟧 4/5	
and1	🟧	🟧 m	🟧 m	
oc	🟧	🟧 *	🟧 *	
tsp4	🟧	⬜	🟩 *	8
carp	🟧	⬜	🟩 m	
hdd11	🟧	⬜	🟩 m	
iunh	🟧	⬜	🟩 *	
pai1	🟧	⬜	🟩 4/5	
hyal4	🟧	⬜	⬜	
ltl	🟧	🟩 m	NA	
caytaxin	🟧	⬜	NA	
and4	🟧	🟧 m	NA	

🟩 Overexpression in eggspot or blotch relative to anal fin tissue
🟧 Underexpression in eggspot or blotch relative to anal fin tissue
⬜ No difference in expression

* <0.05
** <0.01
m = marginally significant < 0.08
NA = no data
4/5 = categorized according to tendency
NA = no statistical support/no tendency

Figure 3 – Gene expression results for the top differentially expressed genes as measured by qPCR **A)** The three species used in our experiments : *A. burtoni*, *P. pulpicans* and *C. macrops*, along with the respective fin pictures, and a scheme defining the different regions of the fin. qPCR was only performed in *P. pulpicans* and *C. macrops*. **B)** *Left panel* shows the results for egg-spot overexpression dataset (table 2). In the first column are the RNAseq results for *A. burtoni*. On the second and third column are the results for *P. pulpican* and *C. macrops* respectively. *Right panel* depicts the results for egg-spot underexpression dataset (table 3). Genes are arranged into clusters by their expression pattern, denoted by numbered vertical black bars. The graphs for each individual qPCR experiment are available in supplementary file 5 (*P. pulpican*) and 6 (*C. macrops*). Details of the statistical analyses used are found in supplementary table 1 (*P.pulpican*) and 2 (*C.macrops*).

Gene expression in *Callochromis macrops*

We measured gene expression of all these genes in another species – *Callochromis macrops.* This species is a member of the ectodine lineage and therefore does not have egg-spots in its anal fin, showing a blotch instead (figure 3A). With this experiment we can test if the same genes that correlate with egg-spot formation are also correlated with blotch morphogenesis.

The gene expression in *C. macrops* was very different from both *A. burtoni* and *P. pulpican* (figure 3B). For dataset A, five genes were also overexpressed in the blotch of *C. macrops*, for dataset B four genes were also underexpressed in the blotch of *C. macrops*. We could not test the expression of five of the genes (for both datasets) because the primers would not amplify at a required efficiency. The ectodine *C.macrops* does not have a genome or transcriptome available and therefore the primers were designed using the haplochromine sequences as a template. Our next step will be to design primers for these genes using the ectodine *Ophthalmotilapia ventralis* transcriptome as a reference (chapter 2). Overall, the results indicate that the two phenotypes might have different genetic basis with some genes in common. These results have to be interpreted with caution due to the lack of statistical power for some of our *C. macrops* gene expression assays. These samples are from wild individuals and not laboratory bred, and therefore they show a higher variation in gene expression measures. As stated above these are ongoing experiments, and we will also add more five individual replicates.

Gene expression of *A. burtoni*, *P. pulpican* and *C. macrops* – Integrative perspective

We clustered the genes according to their expression behaviour in the three species (figure 3B). This grouping can help to address what possible roles these genes play. The groups were specified according to expression patterns in the three species and may represent functionally linked clusters. For the dataset A we defined five groups (1-5) and for the dataset B we defined three (6-8).

Cluster one contains genes that are overexpressed in haplochromines egg-spots and underexpressed in the ectodine blotch. These genes are probably involved in patterning both egg-spot and proximal fin region, since they are more expressed on the anal fin region than on the pigmented region in *C. macrops*. Our hypothesis is that these are part of the egg-spot patterning genes and were co-opted from a fin-patterning pathway. Most of these genes are well-known transcription factors and morphogens (*bmp3b, hoxC12a, hand2*) with patterning/growth functions, or signaling proteins that inhibit the production of certain pigments (*asip1*). Cluster two is a single gene cluster that is overexpressed in the haplochromine egg-spots but does not show a difference in expression between anal fin and blotch of *C. macrops.* This gene is not connected with blotch phenotype or with fin patterning, being most probably only involved in egg-spot morphogenesis. Interestingly, this gene could not be identified and could be a cichlid specific novel, or rapidly evolving, gene. Cluster three are genes that are overexpressed in the egg-spots and blotch of the three species, therefore we think that these genes are related to the production of pigmentation, or they might be patterning the deposition of pigment in all three species. The genes belonging to this cluster are transcription factors (*cecr5*), co-factors (*fhl2a*) [47], cytoskeleton components and kinases (*akap2*). Cluster four is a gene that is overexpressed in *A. burtoni* egg-spots and *C. macrops* blotch, but not on the egg-spots of *P. pulpican*. These two species have in common the fact that the pigments of both blotch and egg-spot are orange, compared to the yellow pigments of *P. pulpican*. This gene (vitronectin [48]) might then be correlated patterning or production of orange pigment, but no role for pigmentation was found in the literature search. Finally, cluster five is composed of genes that are only overexpressed in the egg-spots of *A. burtoni*. These might be fin patterning genes or can even be lineage specific egg-spot genes, participating only on the development and physiology of the trait in this species. This cluster is composed of genes known to participate in fin development (e.g. *rbp7, rbp4, igf1*) [49–51].

We classified dataset B into three clusters (6-8, figure 3B, right panel). Cluster six consisted of genes that were under-expressed in haplochromine egg-spots but showed no difference in expression between tissues in *C. macrops*. These genes seem to correlate only with egg-spot and not with orange pigmentation. Cluster seven genes are underexpressed in all pigmented regions of the fin, either haplochromine egg-spots or ectodini blotch. Finally, cluster eight is composed of genes that are only underexpressed in *A. burtoni* egg-spots. Whilst it is easy to correlate high gene expression with a phenotype, it is tricky to correlate underexpression with the egg-spot phenotype. The most probable explanation is that these genes do not play a role on the morphogenesis of the egg-spot phenotype, participating mainly on the development of the fin and its physiology. There are some interesting patterns though, especially if we consider the function of certain genes. *Axl3* is part of the homeobox gene family, these genes are known for their patterning effects [52]. In this case *axl3* is down-regulated in all three pigmentation phenotypes, meaning that it is overexpressed in the anal fin. Axl3 is actually the gene that shows the highest difference in expression. This gene could be a potential pigmentation inhibitor and therefore should be studied further. Gene clustering according to gene expression is not proof of function, but certainly helps as a hypotheses generator in order to prioritize which genes should be studied first.

4.4 - Discussion

Understanding the genetic and molecular basis of both evolutionary innovation and phenotypic variation is a major challenge in evolutionary biology. A major difficulty has been identifying what is the molecular and developmental basis underlying the novel phenotypes. Using next generation sequencing we present a transcriptional survey of egg-spot tissue in the haplochromine, *Astatotilapia burtoni*. This collection of differentially expressed transcripts represents the biggest dataset for egg-spot candidate genes available and will greatly contribute to the understanding of novel trait emergence, and to the genetic and molecular bases of innovation and variation in the most species rich group of vertebrates – the haplochromine cichlid fishes.

We functionally annotated these datasets and identified enriched functional classes that could generate candidate genes involved in egg-spots. Notably, *Csf1ra* emerged as a candidate, confirming previous work that showed this gene to be involved in the egg-spot development. This shows that our approach is a good candidate gene generator. As we mentioned before, this strategy represents a supervised search for candidates, meaning that we might bias our search our findings towards what is already known and therefore we will find that all the genes involved in the trait are co-opted.

We further studied the expression of 48 genes (dataset A and dataset B) in the egg-spot and blotch other two species – *Pseudotropheus pulpican* (haplochromine) and *Callochromis macrops* (ectodine). Comparing the expression pattern in these three species we defined clusters of genes that could be linked functionally and defined what their potential functions could be. One caveat of our approach is that we can only identify genes that are common to both haplochromine egg-spots. Lineage specific egg-spot genes, which either act in *A. burtoni* egg-spots or in *P. pulpican,* cannot be directly detected with this approach. Cluster five genes could be *A. burtoni* specific egg-spot genes but to confirm if they play a role in the egg-spots we will have to carry out *in situ* hybridization and determine if the gene is expressed in the egg-spots or only on the proximal region of the fin. Nevertheless our clustering approach revealed itself successful. Within each cluster there are some very interesting candidates that should be explored further e.g. *hoxC12a* and *asip1*, but also non-identified genes. It seems that both co-option of known genes and lineage specific genes with new functions are involved in generating diversity. It has been advocated that new traits emerge by the co-option of known conserved regulators [53], but it may well be that these lineage specific genes play a role in trait emergence and diversification, as suggested in [54, 55]. Whether the difference between egg-spot and non-egg-spot lineages resides in coding or regulatory changes, is another topic that should be, and will be, addressed in the near future.

With the gene clustering we cannot prove what the real functions of these genes are, or their exact contribution to phenotypic diversity. What we want to point out is that this type of approach is a very strong and useful hypothesis generator that can help us to choose and characterize genes along with their interaction with phenotypes further. Genes identified through this type of technique can then be tested for signatures of selection in the different cichlid lineages. We can, for example, test if the genes expressed in the egg-spots have a higher dn/ds signature then genes overexpressed in the non-pigmented region of the anal fin. We can test if the different functional categories underwent different selective regimes and try to correlate it with the possible impact of this category on the emergence and diversification of this trait. With the availability of five cichlid genomes and transcriptomes we have the tools for a finer trait mapping. We can perform association mapping, focusing on these genomic regions and correlate it with inter-specific pattern of absence-presence of egg-spots in case we want to address the novelty question, or with the intra-specific variable egg-spot phenotype if we want to address what are the genes that underlie the egg-spot diversity. This task would give us an insight into how pigmentation and patterning pathways behave in an inter- and intra-specific manner. With the availability of the genomes it is also easier to access non-coding sequence surrounding candidate genes, in order to describe possible conserved non-coding elements that might be controlling their expression. Finally, we should study the expression dynamics of these genes whilst the egg-spot is developing, in order to correlate these genes, not only with the adult phenotype, but also with the developmental origin of the trait.

In order to discover how a novelty comes about we need to understand its developmental context, and to address this we need to use comparative methods. Differences in the development of a trait between different species will shed light on the possible function of the underlying genes. Thus, it would be of great interest to broaden the number of species studied representing a range of egg-spot phenotypes. As an example we could test the same types of egg-spot phenotype in different species, including outgroup species that do not show pigmentation in the fin. These experiments could disentangle between lineage specific differences and conserved egg-spot pathways.

As mentioned in the introduction, haplochromine egg-spots are not homologous to *C. macrops* blotch. As expected the gene network underlying the phenotype is different. There is at least one haplochromine species that shows a blotch instead of egg-spots. We do not know if these two traits are homologous (with the blotch representing an intermediate state), or if they have independent origins. This observation has two consequences; first, we can test what makes a blotch and what makes an egg-spot whilst controlling for phylogenetic effect; second we can test for parallel evolution at the genetic level between the two blotch types – haplochromine blotch and ectodine blotch.

As stated above the genetic basis of trait emergence and diversification is still a mystery and much still needs to be done, by establishing a new easily manipulated trait and by generating candidate genes we are moving towards it. Here, we addressed the underlying egg-spot genetics with a transcriptomic approach and defined several solid candidate genes. Studying these genes throughout egg-spot development and in a broader phylogenetic context (~1500 species) will definitely give us some insight into the origin and diversification of this novel trait in the most species rich vertebrate lineage.

4.5 - Conclusions

Using next generation sequencing we described the adult transcriptome of haplochromine egg-spots. This is the first study of its kind demonstrating the power of next generation sequencing to generate candidate genes, especially for traits where little is known about the underlying genetics. We further tested and verified the expression of a sub-set of these genes in another haplochromine species, identifying a set of egg-spot genes that will serve as useful resource for future research on the genetics and evolution of this trait. Finally, we provide evidence that both co-option of old genes and lineage specific genes are involved in the formation of the egg-spot phenotype, and that therefore suggest both mechanisms play a role in the evolution of novel traits. Combining the study of these genes in egg-spot development with comparisons across the diverse range of naturally occurring haplochromine cichlid species will definitely lead to major advances in the field of emergence and diversification of novel traits.

4.6 - Methods

Samples

Astatotilapia burtoni and *Pseudotropheus pulpican* bred laboratory strains are kept at the University of Basel (Switzerland) under standard conditions (12h light/12h dark; 26°C, pH7). *Callochromis macrops* individuals were captured in lake Tanganyika, Mpulungu (Zambia). Dissections were carried out *in situ*, the tissues were stored in RNAlater (Ambion, USA) and shipped to the University of Basel. Before dissection all individual euthanized with MS222 (Sigma-Aldrich, USA) before dissection following approved procedures (permit nr. 2317 issued by the cantonal veterinary office).

RNA extractions

All RNA extractions and cDNA production were performed as described in chapter three.

Differential gene expression analysis using RNAseq – Illumina

The anal fins from six *Astatotilapia burtoni* male juveniles were dissected and RNA was extracted from egg-spot tissue and anal fin tissue for each individual (figure 1A). One microgram of RNA per sample was sent for library construction and Illumina sequencing at the Department of Biosystems Science and Engineering (D-BSSE), University of Basel and ETH Zurich. Samples were run in two lanes of an Illumina Genome Analyzer IIx (maximum read length was 50 bp). Individuals one to three were sequenced in one lane, and the other three individuals in a second lane. Anal fin samples and egg-spot samples were tagged in order to differentiate between them.

The reads were mapped against a reference transcriptome, which consisted of a pool of *A. burtoni* embryo transcripts sequenced at The Broad Institute (Massachusetts, USA) (http://www.broadinstitute.org/models/tilapia). This transcriptome reference was indexed with NOVOINDEX (www.novocraft.com) using default parameters. Using NOVOALIGN (www.novocraft.com), we mapped the reads from each library against the reference transcriptome using default parameter except for the following: a) maximum alignment score (t) of 30; b) a minimum of good quality base pairs per read (l) of 25; c) a successive trimming factor (s) of five. Reads that did not match these criteria were discarded. The reference transcriptome is a redundant database, containing several isoforms of the same gene. For this reason all read alignment locations were reported in SAM output format (rALL, oSAM). The

SAM file was then transformed into a count file (number of reads per transcript) using SAMTOOLS version 0.1.18 [56]. The countfiles from each sample was then concatenated into one single dataset and analyzed with the Bioconductor R package EdgeR [57]. We tested for differential expression between egg-spot and anal fin samples, using anal fin as reference. Since the samples were paired we included the individual information in the statistical model. We used a negative binomial GLM based on common dispersion using the individual as the blocking factor, i.e. we tested for differences in expression between egg-spot and anal fin within individuals. Transcripts were considered as differentially expressed if, after the correction for multiple testing, the false discovery rate (FDR) was lower than 0.01 [58].

Annotation

Gene ontology (GO) [59] annotation of the differential expressed transcripts was conducted with Blast2GO version 2.5.0 [36]. BLASTx searches were done against the non-redundant database (nr) using the QBLAST for multiple queries, setting the *e*-value to 1.0 x 10^{-6}, the high scoring segment length cut off greater than 33, and the number of hits to 10. InterproScan annotation was also retrieved and combined with the blast2GO annotation [60]. We also assigned to each transcript an enzyme commission (EC) number and metabolic pathways using Kyoto encyclopedia of genes and genomes (KEGG) database [61]. These GO terms were used to estimate transcript function. The table with the list of the differential expressed transcripts their respective values of expression and the GO terms are provided as supplementary material (supplementary file 1). Overall differences in level two GO term distributions between the two datasets were tested used chi-squared tests. Differences, between the datasets, in proportion of genes for individual level two GO terms were tested using chi-squared tests with *p*-values adjusted for multiple tests using Bonferroni corrections [62].

Differential GO term representation

We tested for differential GO composition between overexpressed and underexpressed egg-spot transcripts. For this analysis we merged the isoforms to their common gene, because the number of isofoms would influence the quantification of specific GO terms. Each transcript has a code where the first term codes for the parent contig and the third term codes for the alternatively spliced transcripts, with each code separated by an underscore (CompX_cX_seqX). Using in-house Python scripts we merged the GO terms from all the transcripts GO terms to their parent gene (supplementary file 2). We then converted the file into an annotation file readable by blast2GO. From then on the analysis was done using only the gene information. We tested for differential GO term representation between the two datasets with a Fisher's exact test. A GO term was considered differentially represented if FDR < 0.05. The list of differentially represented terms is provided as supplementary material (supplementary file 3).

Gene expression analysis using qPCR

The expression of 48 genes (24 most overexpressed genes in the egg-spot region and 24 most underexpressed genes in the egg-spot) was further studied in two other species - *Pseudotropheus pulpican* and *Callochromis macrops*. For one gene in the overexpressed dataset, there was no reference genome sequence available, and this was omitted to avoid working with assembly errors. The 25th most differentially expressed gene was used instead. Primers were tested in both species and in cases where the primers pair did not work for both species we designed new primers for *C.macrops*. Primer design was performed using the GenScript Real-time PCR (TaqMan) Primer Design software available at www.genscript.com/ssl-bin/app/primer. When possible primers were designed in exon spanning regions to avoid gDNA contamination. Genes studied and primer sequences are available in supplementary material (supplementary file 4). For each species five individuals were used. **qPCR experiment 1**: Gene expression was compared between the non-egg-spot anal fin tissue and the egg-spot tissue of *P.pulpican.* This species has its egg-spot in a different position in the fin compared to *A.burtoni* (figure 3A). **qPCR experiment 2**: Gene expression was compared between the non-blotch anal fin tissue and blotch tissue of *C.macrops* (figure 3A).

The reactions were run using the StepOnePlus™ Real-Time PCR system (Applied Biosystems, USA) using SYBR Green master mix (Roche, Switzerland) following the manufacturer's protocols. All reactions were performed with an annealing temperature of 58ºC using a final concentration of cDNA of 1ng/μl and a final primer concentration of 200ng/μl. The comparative threshold cycle (CT) method [63] was used to calculate the relative concentrations between tissues, where anal fin was used as the reference tissue and Ribosomal protein L7 (*rpl7*) and the Ribosomal protein SA3 (*rpsa3*) genes as endogenous controls. Primer efficiencies were calculated using standard curves. Efficiency values of test primers are comparable to the efficiency of endogenous control primers (*rpl7, rspa3*) and are available in supplementary file 4.

Significant differential gene expression between egg-spot/blotch and anal fin was tested with a paired *t*-test, or a Mann-Whitney non-parametric test if data did not conform to the assumptions of a *t*-test. Statistics were carried out using GRAPHPAD Prism version 5.0a for Mac OS X (www.graphpad.com). The details of the statistical results are given in supplementary table 1 (*P. pulpican*) and supplementary table 2 (*C. macrops*). The differences in expression were categorized and plotted in figure 3. Due to our small sample size some of the results were not significant, in this cases we categorized the data according to tendencies, if four out of five individuals showed difference in expression we would categorize them accordingly to over- or underexpression tendency. Individual graphs for each gene studied are available in supplementary file 5 (*P. pulpican*) and supplementary file 6 (*C. macrops*).

4.7 - Supplementary Information

Supplementary Table 1 Test statistics for qPCR analysis of *P. pulpicans* genes in egg-spots

Gene Name	Cluster	Test used	Test Statistics: t, df for t-test, or Sum of ranks for MW	Nos of pairs for *t*-test or Mann Whitney U Stat	*p*-value
akap2	3	Mann-Whitney	10 , 35	0	0.0159
and1	7	Paired t-test	t=2.462 df=4	5	0.0696
and4	no cluster	Mann-Whitney	28 , 17	2	0.0635
asip1	1	Paired t-test	t=2.045 df=4	5	0.1104
axl3	7	Mann-Whitney	30 , 15	0	0.0159
bmp3b	1	Paired t-test	t=4.538 df=4	5	0.0105
carp	8	Mann-Whitney	23 , 32	8	0.4206
caytaxin	no cluster	Mann-Whitney	28 , 27	12	1
cd81	6	Mann-Whitney	30 , 15	0	0.0159
cecr5	3	Mann-Whitney	15 , 40	0	0.0079
col9a1	6	Mann-Whitney	37 , 18	3	0.0556
comp116662_c0	6	Mann-Whitney	40 , 15	0	0.0079
comp17910_c0	3	Mann-Whitney	15 , 40	0	0.0079
comp23328_c0	2	Mann-Whitney	15 , 40		0.0079
comp23699_c0	no cluster	NA	NA	NA	NA
comp24816_c0	1	Mann-Whitney	15 , 40	0	0.0079
comp29158_c0	6	Mann-Whitney	30 , 15	0	0.0159
comp4443_c1	no cluster	Mann-Whitney	21 , 34	6	0.2222
cytl1	1	Paired t-test	t=3.548 df=4	5	0.0238
fhl2a	3	Paired t-test	t=5.031 df=4	5	0.0073
fmdo	5	Mann-Whitney	39 , 16	1	0.0159
hand2	1	Paired t-test	t=6.043 df=4	5	0.0038
hbaa	6	Mann-Whitney	28 , 17	2	0.0635
hbba	6	Mann-Whitney	40 , 15	0	0.0079
hdd11	8	Paired t-test	t=1.454 df=3	4	0.242
HoxC12a	1	Paired t-test	t=4.720 df=3	4	0.018
hyal4	no cluster	Mann-Whitney	25 , 30	10	0.6905
ifON3	5	Mann-Whitney	22 , 33	7	0.3095
igf1	5	Paired t-test	t=2.387 df=4	5	0.0754
IgSF10	5	Paired t-test	t=9.751 df=3	4	0.0023
iunh	8	Mann-Whitney	21 , 34	6	0.2222
LOC100695447	6	Paired t-test	t=2.869 df=3	4	0.0641
LOC100708826	1	Mann-Whitney	15 , 40	0	0.0079
loxl4	7	Mann-Whitney	29 , 16	1	0.0317
ltl	no cluster	Paired t-test	t=2.698 df=4	5	0.0542
matn4	6	Mann-Whitney	30 , 15	0	0.0159
mmp13	6	Mann-Whitney	30 , 15	0	0.0159
mucin2	no cluster	NA	NA	NA	NA
oc	7	Mann-Whitney	30 , 15	0	0.0159
pai1	8	Mann-Whitney	24 , 21	6	0.4127
phospho1	6	Mann-Whitney	30 , 15	0	0.0159
rbp4a	5	Paired t-test	t=1.269 df=4	5	0.2732

Supplementary table 1 continued

rbp7	5	Mann-Whitney	40 , 15	0	0.0079
sfr5	5	Paired t-test	t=6.755 df=4	5	0.0025
slc13m5	6	Mann-Whitney	30 , 15	0	0.0159
tsp4	8	Mann-Whitney	18 , 27	8	0.7302
vtn	4	Mann-Whitney	39 , 16	1	0.0159
zygin1	no cluster	Mann-Whitney	10 , 35	0	0.0159

Supplementary Table 2 Test statistics for qPCR analysis of *C.macrops* genes in blotches

Gene Name	Cluster	Test used	Test Statistics: t, df for t-test, or Sum of ranks for MW	Nos of pairs for *t*-test or Mann Whitney U Stat	*p*-value
akap2	3	paired t-tested	t=4.058 df=4	5	0.0154
and1	7	paired t-tested	t=2.507 df=4	5	0.0663
and4	no cluster	NA	NA	NA	NA
asip1	1	paired t-tested	t=1.259 df=4	5	0.2765
axl3	7	Mann-Whitney	35 , 20	5	0.1508
bmp3b	1	Mann-Whitney	32 , 23	8	0.4206
carp	8	paired t-tested	t=2.733 df=3	4	0.0718
caytaxin	no cluster	paired t-tested	NA	NA	NA
cd81	6	Mann-Whitney	32 , 23	8	0.4206
cecr5	3	Mann-Whitney	15 , 40	0	0.0079
col9a1	6	Mann-Whitney	29 , 26	11	0.8413
comp116662_c0	6	Mann-Whitney	19 , 36	4	0.0952
comp17910_c0	3	Mann-Whitney	26 , 29	11	0.8413
comp23328_c0	2	paired t-tested	t=0.7693 df=4	5	0.4846
comp23699_c0	no cluster	Mann-Whitney	23 , 22	8	0.7302
comp24816_c0	1	paired t-tested	t=1.533 df=4	5	0.2
comp29158_c0	6	Mann-Whitney	25 , 30	10	0.6905
comp4443_c1	no cluster	paired t-tested	NA	NA	NA
cytl1	1	paired t-tested	t=3.355 df=4	5	0.0284
fhl2a	3	Mann-Whitney	16 , 39	1	0.0159
fmdo	5	Mann-Whitney	29 , 26	11	0.8413
hand2	1	Mann-Whitney	35 , 20	5	0.1508
hbaa	6	Mann-Whitney	26 , 29	11	0.8413
hbba	6	Mann-Whitney	29 , 26	11	0.8413
hdd11	8	Mann-Whitney	18 , 37	3	0.0556
HoxC12a	1	Mann-Whitney	33 , 22	7	0.3095
hyal4	no cluster	Mann-Whitney	32 , 23	8	0.4206
ifON3	5	paired t-tested	t=3.278 df=4	5	0.0306
igf1	5	Mann-Whitney	36 , 19	4	0.0952
IgSF10	5	Mann-Whitney	35 , 20	5	0.1508
iunh	8	paired t-tested	t=2.805 df=4	5	0.0486
LOC100695447	6	Mann-Whitney	31,14	4	0.1905
LOC100708826	1	paired t-tested	t=1.759 df=4	5	0.1534
loxl4	7	paired t-tested	t=2.949 df=4	5	0.042
ltl	no cluster	paired t-tested	NA	NA	NA
matn4	6	Mann-Whitney	30 , 25	10	0.6905
mmp13	6	Mann-Whitney	29 , 26	11	0.8413
mucin2	no cluster	Mann-Whitney	25 , 30	10	0.6905
oc	7	Mann-Whitney	38 , 17	2	0.0317
pai1	8	paired t-tested	t=2.039 df=4	5	0.1111

Supplementary table 2 continued

phospho1	6	Mann-Whitney	30 , 25	10	0.6905
rbp4a	5	Mann-Whitney	40 , 15	0	0.0079
rbp7	5	Mann-Whitney	40 , 15	0	0.0079
sfr5	5	Mann-Whitney	31 , 24	9	0.5476
slc13m5	6	Mann-Whitney	31 , 14	4	0.1905
tsp4	8	paired t-tested	t=2.825 df=4	5	0.0476
vtn	4	Mann-Whitney	15 , 40	0	0.0079
zygin1	no cluster	paired t-tested	NA	NA	NA

Supplementary material provided in digital format

Supplementary file 1 Differentially expressed transcripts with respective expression values and GO annotation

Supplementary file 2 Differentially expressed contigs with respective GO annotation

Supplementary file 3 List of differentially represented GO terms between overexpressed and underexpressed datasets

Supplementary file 4 List of primers used in this study (includes sequence and PCR efficiency values for each

Supplementary file 5 *P. pulpican* graphs for qPCR experiments for individual genes gene

Supplementary file 6 *C. macrops* graphs for qPCR experiments for individual genes gene

4.8 - Acknowledgments

We would like to thank Anya Theis for the fish photographs. This study was supported by a doctoral research fellowship from the Fundação para a Ciência e Tecnologia (SFRH/BD/43421/2008) and the European Research Council (Starting Grant 'INTERGENADAPT').

4.9 - References

1. Wagner GP, Lynch VJ: **Evolutionary novelties.** *Current Biology* 2010, **20**:R48-52.

2. Moczek AP: **On the origins of novelty in development and evolution.** *BioEssays* 2008, **30**:432-447.

3. Pigliucci M: **What, if Anything, Is an Evolutionary Novelty?** *Philosophy of Science* 2008, **75**:887-898.

4. True JR, Carroll SB: **Gene co-option in physiological and morphological evolution.** *Annual Review of Cell and Developmental Biology* 2002, **18**:53-80.

5. Prud'homme B, Gompel N, Carroll SB: **Emerging principles of regulatory evolution.** *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**:8605-8612.

6. Brayer KJ, Lynch VJ, Wagner GP: **Evolution of a derived protein–protein interaction between HoxA11 and Foxo1a in mammals caused by changes in intramolecular regulation**. *Proceedings of the National Academy of Sciences of the United States of America* 2011, **108**:E414–E420.

7. Martin A, Reed RD: **Wingless and aristaless2 define a developmental ground plan for moth and butterfly wing pattern evolution.** *Molecular Biology and Evolution* 2010, **27**:2864-2878.

8. Milde S, Hemmrich G, Anton-Erxleben F, Khalturin K, Wittlieb J, Bosch TCG: **Characterization of taxonomically restricted genes in a phylum-restricted cell type.** *Genome Biology* 2009, **10**:R8.

9. Knox K, Baker JC: **Genomic evolution of the placenta using co-option and duplication and divergence.** *Genome Research* 2008, **18**:695-705.

10. Carroll SB: **Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution.** *Cell* 2008, **134**:25-36.

11. Hoekstra HE, Coyne JA: **The locus of evolution: evo devo and the genetics of adaptation.** *Evolution* 2007, **61**:995-1016.

12. Lynch VJ, Wagner GP: **Resurrecting the role of transcription factor change in developmental evolution.** *Evolution* 2008, **62**:2131-2154.

13. Stern D, Orgogozo V: **Is Genetic Evolution Predictable?** *Science* 2009, **323**:746-751.

14. Monteiro A, Podlaha O: **Wings, horns, and butterfly eyespots: how do complex traits evolve?** *PLoS Biology* 2009, **7**:e37.

15. Endler J: **Natural selection on color patterns in Poecilia reticulata**. *Evolution* 1980, **34**:76-91.

16. Hubbard JK, Uy JAC, Hauber ME, Hoekstra HE, Safran RJ: **Vertebrate pigmentation: from underlying genes to adaptive function.** *Trends in Genetics* 2010, **26**:231-239.

17. Hoekstra HE: **Genetics, development and evolution of adaptive pigmentation in vertebrates.** *Heredity* 2006, **97**:222-234.

18. Wittkopp PJ, Beldade P: **Development and evolution of insect pigmentation: genetic mechanisms and the potential consequences of pleiotropy.** *Seminars in Cell & Developmental Biology* 2009, **20**:65-71.

19. Kocher TD: **Adaptive evolution and explosive speciation: the cichlid fish model.** *Nature Reviews Genetics* 2004, **5**:288-298.

20. Seehausen O: **African cichlid fish: a model system in adaptive radiation research.** *Proceedings of the Royal Society B* 2006, **273**:1987-1998.

21. Salzburger W: **The interaction of sexually and naturally selected traits in the adaptive radiations of cichlid fishes.** *Molecular Ecology* 2009, **18**:169-185.

22. Salzburger W, Mack T, Verheyen E, Meyer A: **Out of Tanganyika: genesis, explosive speciation, key-innovations and phylogeography of the haplochromine cichlid fishes.** *BMC Evolutionary Biology* 2005, **5**:17.

23. Hert E: **The function of egg-spots in an African mouth-brooding cichlid fish**. *Animal Behaviour* 1989, **37**:726–732.

24. Hert E: **Female choice based on egg-spots in Pseudotropheus aurora Burgess 1976, a rock-dwelling cichlid of Lake Malawi, Africa.** *Journal of Fish Biology* 1991, **38**:951–953.

25. Lehtonen TK, Meyer A: **Heritability and adaptive significance of the number of egg-dummies in the cichlid fish Astatotilapia burtoni.** *Proceedings of the Royal Society B* 2011, **278**:2318-2324.

26. Theis A, Salzburger W, Egger B: **The function of anal fin egg-spots in the cichlid fish Astatotilapia burtoni.** *PloS One* 2012, **7**:e29878.

27. Salzburger W, Braasch I, Meyer A: **Adaptive sequence evolution in a color gene involved in the formation of the characteristic egg-dummies of male haplochromine cichlid fishes.** *BMC Biology* 2007, **5**:51.

28. Fryer G, Iles T: *The Cichlid Fishes of the Great Lakes of Africa: Their Biology and Evolution*. Edinburgh, UK: Oliver & Boyd; 1972.

29. Fujii R: **The regulation of motile activity in fish chromatophores.** *Pigment Cell Research* 2000, **13**:300-319.

30. Kelsh RN: **Review : Pigment Gene Focus Genetics and Evolution of Pigment Patterns in Fish**. *Cell Research* 2004, **17**:326-336.

31. Terai Y, Morikawa N, Kawakami K, Okada N: **The complexity of alternative splicing of hagoromo mRNAs is increased in an explosively speciated lineage**

in East African cichlids. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**:12798-803.

32. Terai Y, Morikawa N, Kawakami K, Okada N: **Accelerated Evolution of the Surface Amino Acids in the WD-Repeat Domain Encoded by the hagoromo Gene in an Explosively Speciated Lineage of East African Cichlid**. *Molecular Biology and Evolution* 2002, **19**:574-578.

33. Roberts RB, Ser JR, Kocher TD: **Sexual conflict resolved by invasion of a novel sex determiner in Lake Malawi cichlid fishes.** *Science* 2009, **326**:998-1001.

34. Ungerer MC, Johnson LC, Herman M a: **Ecological genomics: understanding gene and genome function in the natural environment.** *Heredity* 2008, **100**:178-83.

35. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics.** *Nature reviews. Genetics* 2009, **10**:57-63.

36. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.** *Bioinformatics (Oxford, England)* 2005, **21**:3674-6.

37. Kelsh RN, Harris ML, Colanesi S, Erickson C a: **Stripes and belly-spots -- a review of pigment cell morphogenesis in vertebrates.** *Seminars in cell & developmental biology* 2009, **20**:90-104.

38. Dupin E, Sommer L: **Neural crest progenitors and stem cells: from early development to adulthood.** *Developmental biology* 2012, **366**:83-95.

39. Christiansen JH, Coles EG, Wilkinson DG: **Molecular control of neural crest formation, migration and differentiation.** *Current opinion in cell biology* 2000, **12**:719-24.

40. Braasch I, Liedtke D, Volff J-N, Schartl M: **Pigmentary function and evolution of tyrp1 gene duplicates in fish.** *Pigment cell & melanoma research* 2009, **22**:839-50.

41. Parichy DM, Rawls JF, Pratt SJ, Whitfield TT, Johnson SL: **Zebrafish sparse corresponds to an orthologue of c-kit and is required for the morphogenesis of a subpopulation of melanocytes, but is not essential for hematopoiesis or primordial germ cell development.** *Development (Cambridge, England)* 1999, **126**:3425-36.

42. Parichy DM: **Temporal and cellular requirements for Fms signaling during zebrafish adult pigment pattern development**. *Development* 2003, **130**:817-833.

43. Pick L, Heffer A: **Hox gene evolution: multiple mechanisms contributing to evolutionary novelties.** *Annals of the New York Academy of Sciences* 2012:1-18.

44. Yelon D, Ticho B, Halpern ME, Ruvinsky I, Ho RK, Silver LM, Stainier DY: **The bHLH transcription factor hand2 plays parallel roles in zebrafish heart and pectoral fin development.** *Development* 2000, **127**:2573-82.

45. Gamer LW, Ho V, Cox K, Rosen V: **Expression and function of BMP3 during chick limb development.** *Developmental dynamics* 2008, **237**:1691-8.

46. Manceau M, Domingues VS, Mallarino R, Hoekstra HE: **The developmental role of Agouti in color pattern evolution.** *Science (New York, N.Y.)* 2011, **331**:1062-5.

47. Johannessen M, Møller S, Hansen T, Moens U, Van Ghelue M: **The multifunctional roles of the four-and-a-half-LIM only protein FHL2.** *Cellular and Molecular Life Sciences* 2006, **63**:268-284.

48. Felding-Habermann B, Cheresh D a: **Vitronectin and its receptors.** *Current opinion in cell biology* 1993, **5**:864-8.

49. Blum N, Begemann G: **Retinoic acid signaling controls the formation, proliferation and survival of the blastema during adult zebrafish fin regeneration.** *Development* 2012, **139**:107-16.

50. Tingaud-Sequeira A, Forgue J, André M, Babin PJ: **Epidermal transient down-regulation of retinol-binding protein 4 and mirror expression of apolipoprotein Eb and estrogen receptor 2a during zebrafish fin and scale development.** *Developmental dynamics* 2006, **235**:3071-9.

51. Chablais F, Jazwinska A: **IGF signaling between blastema and wound epidermis is required for fin regeneration.** *Development* 2010, **137**:871-9.

52. McGonnell IM, Graham A, Richardson J, Fish JL, Depew MJ, Dee CT, Holland PWH, Takahashi T: **Evolution of the Alx homeobox gene family: parallel retention and independent loss of the vertebrate Alx3 gene.** *Evolution & development* 2011, **13**:343-51.

53. Carroll SB: **Evolution at two levels: on genes and form.** *PLoS biology* 2005, **3**:e245.

54. Nowick K, Stubbs L: **Lineage-specific transcription factors and the evolution of gene regulatory networks.** *Briefings in Functional Genomics* 2010, **9**:65-78.

55. Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TCG: **More than just orphans: are taxonomically-restricted genes important in evolution?** *Trends in Genetics* 2009, **25**:404-13.

56. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25**:2078-2079.

57. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics* 2010, **26**:139-140.

58. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing**. *Journal of the Royal Statistical Society. Series B (Methodological)* 1995, **57**:289-300.

59. Ashburner M, Ball C, Blake J: **Gene Ontology: tool for the unification of biology**. *Nature Genetics* 2000, **25**:25-29.

60. Zdobnov E, Apweiler R: **InterProScan – an integration platform for the signature-recognition methods in InterPro**. *Bioinformatics* 2001, **17**:847-848.

61. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes.** *Nucleic acids research* 2000, **28**:27-30.

62. Bonferroni CE: **Teoria statistica delle classi e calcolo delle probabilità**. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze* 1936, **8**:62.

63. Pfaffl MW: **A new mathematical model for relative quantification in real-time RT-PCR.** *Nucleic Acids Research* 2001, **29**:e45.

Chapter 5

How cichlids diversify

M. Emília Santos, Walter Salzburger

Perspective published in Science

The diverse assemblages of cichlid fishes in East Africa are the ideal model system for an integrative survey of how and why organisms diversify

at older ages can be accounted for by appropriate allocation models (*10*). In such models, the force of selection declines with age, but though important, this decline is not decisive in molding fertility and mortality patterns.

What is decisive is the "option set" of a species, which can be summarized by the feasible combinations of survival and reproduction at all ages over the life span. Option sets differ widely: For some species, extra investment in repair and maintenance substantially reduces fertility; for other species there is little impact; for yet other species enhanced repair and maintenance decrease current but increase future fecundity. The details of such option sets shape age patterns of growth, fertility, and mortality (*8*, *11*).

Little is known about what types of constraints favor a pattern of aging with increasing mortality and decreasing fertility (senescent) versus alternative patterns with constant or declining mortality and constant or increasing fertility (nonsenescent). Life-history models suggest that the marginal costs and benefits of energy allocation play a central role (*8*, *11*). To test this and to explore other hypotheses, it would be informative to compare plants, for which growth and reproduction flexibly adapt to environmental conditions (*12*), to animals, for which growth and reproduction are more rigid and distinct (*8*). In contrast to vertebrates, plants capable of vegetative reproduction can create offspring by splitting off body parts. Thereby an investment in growth effectively becomes an investment in reproduction. Species that are small but long-lived (such as hydra in the laboratory), that can reproduce either sexually or asexually (such as daphnia), or that face highly uncertain environments [such as desert plants (*12*)] may also be good candidates for studies of how allocation options shape patterns of aging.

Research on the evolution of aging should focus on unraveling those differences in species' option sets that lead to senescent versus nonsenescent aging patterns. A major barrier in accomplishing this has been the lack of laboratory, zoo, and field evidence about age patterns of growth, maintenance, fertility, and mortality for species across the tree of life. New statistical methods and software now permit the extraction of mortality patterns from field data that are sporadic or are missing observations (*13*). Further development of life-history models hinges on more extensive and reliable data as well as on experiments to reveal how much allocation of additional resources to, say, faster growth or a more effective immune system affects lifetime fertility and survival. Fundamental understanding of why humans deteriorate so sharply (*14*) compared with other species, why human mortality has fallen so dramatically (*15*), and whether aging can be further delayed or even slowed (*16*) depends on knowledge of why some species senesce and others do not.

References and Notes

1. T. B. L. Kirkwood, *Nature* **270**, 301 (1977).
2. P. B. Medawar, in *Uniqueness of the Individual* (Lewis, London, 1952), pp. 44–70.
3. G. C. Williams, *Evolution* **11**, 398 (1957).
4. W. D. Hamilton, *J. Theor. Biol.* **12**, 12 (1966).
5. O. R. Jones *et al.*, *Ecol. Lett.* **11**, 664 (2008).
6. S. C. Stearns, *The Evolution of Life Histories* (Oxford Univ. Press, Oxford, New York, 1992).
7. L. Partridge, R. Sibly, R. J. H. Beverton, W. G. Hill, *Philos. Trans. Biol. Sci.* **332**, 3 (1991).
8. A. Baudisch, *Inevitable Senescence? Contributions to Evolutionary Demographic Theory*, Demographic Research Monographs (Springer, Berlin, 2008).
9. J. W. Vaupel *et al.*, *Theor. Popul. Biol.* **65**, 339 (2004).
10. M. J. Dańko *et al.*, *PLoS ONE* **7**, e34146 (2012).
11. A. Baudisch, *Gerontology* 10.1159/000341861 (2012).
12. R. Salguero-Gómez, B. C. Casper, *J. Ecol.* **98**, 312 (2010).
13. F. Colchero *et al.*, *Methods Ecol. Evol.* **3**, 466 (2012).
14. A. Baudisch, *Methods Ecol. Evol* **2**, 375 (2011).
15. O. Burger, A. Baudisch, J. W. Vaupel, *Proc. Natl. Acad. Sci. U.S.A.* 10.1073/pnas.1215627109 (2012).
16. J. W. Vaupel, *Nature* **464**, 536 (2010).
17. F. B. Turner, K. H. Berry, D. C. Randall, G. C. White, "Population ecology of the desert tortoise at Goffs, California, 1983–1986. Report No. 87-RD-81" (Southern California Edison Company, 1987).
18. D. E. Martínez, *Exp. Gerontol.* **33**, 217 (1998).

Acknowledgments: We thank L. Partridge, F. Colchero, D. Conde, D. Levitis, O. Jones, R. Salguero-Gomez, A. Scheuerlein, and the Evodemo group at the Max Planck Institute for Demographic Research. Supported by NIH grant AG-031719.

10.1126/science.1226467

EVOLUTION

How Cichlids Diversify

The extreme diversity of cichlid fishes in East Africa helps to elucidate how and why organisms diversify.

M. Emília Santos and Walter Salzburger

How is genetic variation connected to morphological evolution? How did Earth's spectacular organismal diversity evolve and how is it maintained? To answer these fundamental questions, scientists must understand how organisms function and diversify and how they interact with other organisms and the environment. Recent studies of cichlids, including (*1*–*7*), are beginning to provide insights into the basis of diversification in this exceptionally diverse fish family.

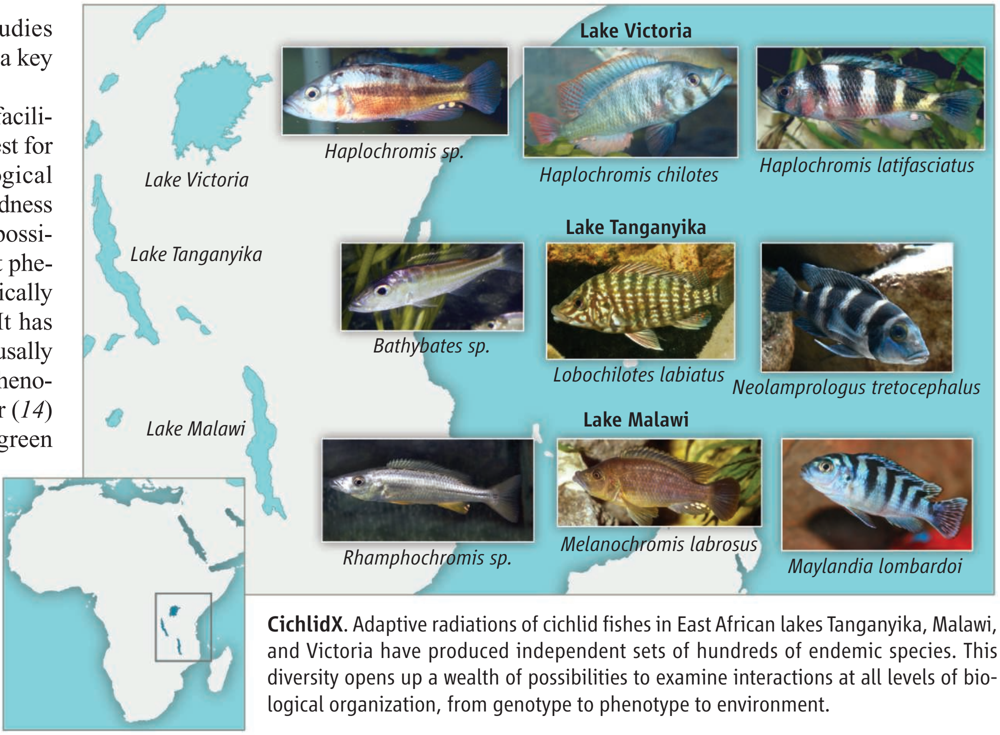
Many widely used biological model systems only provide limited insights into organismal diversification. Traditional laboratory-based model organisms tell us little about how organisms survive, adapt, behave, and reproduce in the wild. Model organisms used in evolutionary and ecological research, on the other hand, are often difficult to breed, their genomes are poorly characterized, and few genetic and developmental tools are available to study them. Furthermore, most established model systems are not very diverse taxonomically and phenotypically. Notable exceptions are instances of adaptive radiation, that is, the rapid origination of a multitude of phenotypically diverse species from a common ancestor through adaptation to distinct ecological niches (*8*, *9*). Famous examples of adaptive radiations include Darwin's finches on the Galápagos archipelago, silversword plants on Hawaii, anole lizards on islands of the Caribbean, and cichlid fishes in East Africa.

In the case of cichlids, hundreds of endemic species evolved independently in each of the three East African Great Lakes: Victoria, Malawi, and Tanganyika. Cichlids thus form by far the most species-rich extant adaptive radiations. They split up into distinct species in such little time that their DNA is still almost identical, a situation comparable to an experimental mutagenesis screen, yet in a natural environment (*10*).

Analyses of draft genome and transcriptome sequences have demonstrated the potential provided by such data (*1*, 2, *5*, *7*, *11*). Loh *et al.* (*1*), for example, investigated microRNA genes, which are important agents for the regulation of gene expression, and detected signatures of divergent natural selection in microRNA target sites among Lake Malawi cichlids. A comparative transcriptome analysis revealed little divergence at protein-coding sequences but high diversity in untranslated regions that are impor-

Zoological Institute, University of Basel, Vesalgasse 1, CH-4051 Basel, Switzerland. E-mail: emilia.santos@unibas.ch; walter.salzburger@unibas.ch

tant for gene regulation (*2*). These studies suggest that regulatory evolution plays a key role in cichlid diversification.



CichlidX. Adaptive radiations of cichlid fishes in East African lakes Tanganyika, Malawi, and Victoria have produced independent sets of hundreds of endemic species. This diversity opens up a wealth of possibilities to examine interactions at all levels of biological organization, from genotype to phenotype to environment.

Draft genome sequences have also facilitated developmental studies and the quest for genes underlying adaptive morphological traits (*11*). Because of their close relatedness and amenability to aquarium life, it is possible to cross cichlid species with distinct phenotypes in the laboratory to then genetically map key evolutionary traits (*12*, *13*). It has now become possible to directly and causally link molecules to phenotypes through phenotypic engineering. Fujimura and Kocher (*14*) created transgenic tilapia that express green fluorescent protein under the control of a *Xenopus* promoter. This method allows the function of genes in cichlids to be studied directly.

A wealth of information, spanning decades of research (*15*, *16*), is available on the evolution, ecology, morphology, and behavior for many cichlid species and communities. These diverse data open up various possibilities to examine the relative importance of natural and sexual selection, contingency, and determinism to cichlid evolution and to observe the evolution of fitness-relevant traits as well as their underlying genes in action.

For example, Seehausen *et al*. (*17*) and Miyagi *et al*. (*3*) have examined the role of visual pigments in the recent divergence of Lake Victoria cichlids. The heterogeneous light conditions in this lake led to diversifying selection on *opsin* genes as a function of water depth. The divergence in opsins, in turn, affects sexual selection, because differences in color perception influence the female preference for male coloration (*17*). Here, the interplay between natural and sexual selection resulted in speciation in the absence of geographic barriers through selection on a sensory system ("sensory drive").

In other cases, natural and sexual selection act in opposite directions. An orange-blotch coloration is common among females of Lake Malawi cichlids and provides camouflage over boulders. Blotched males, on the other hand, seem to have a selective disadvantage because they do not possess the nuptial coloration that attracts females. Roberts *et al*. have recently shown (*12*) how this conflict between natural selection (the orange blotch pattern provides camouflage) and sexual selection (orange blotch males are less likely to reproduce) is resolved. A new female sex-determining gene has evolved in linkage with the *pax7* gene that makes the orange blotch coloration. This linkage leads to low recombination; therefore, mostly females have this coloration.

Perhaps the most important feature of cichlid adaptive radiations, at least in the context of speciation, is that they come in replicates, because lakes Malawi, Victoria, and Tanganyika each have their own cichlid assemblage (see the figure). "Nature's grand experiment in evolution" (*16*) therefore provides an opportunity for comparing patterns and processes of diversification—especially because both very species-rich (radiating) and species-poor (nonradiating) groups of cichlids exist.

In a recent analysis focusing on 46 African lakes (*4*), Wagner *et al*. concluded that cichlids are more prone to radiate if they are sexually dichromatic (with males and females showing different pigmentation patterns), live in deeper and older lakes, and occupy regions with more solar energy input. The combination of environmental conditions and sexual dichromatism does not explain all cichlid radiations; for example, there are no differences in coloration between males and females in the ~100 species of lamprologines in Lake Tanganyika. Nevertheless, Wagner *et al*. demonstrate that patterns of diversification can at least partially be predicted.

The main outcome of "evolution in replicates" is a high abundance of convergent phenotypes, which are perfectly suited to elucidate the molecular mechanisms and/or developmental constraints involved in parallel evolution. Colombo *et al*. (*5*), for example, identified striking similarities in the genetics underlying the thick-lipped phenotype found in East African and Central American cichlid radiations, which are separated by almost 100 million years of independent evolution. That phenotypic parallelism is not restricted to morphology in cichlids is, for example, highlighted by the repeated transition of parental care strategies in the Ectodini, a group of mouthbrooding cichlids from Lake Tanganyika (*6*), illustrating once more the broad scope of traits and topics that can be tackled with the cichlid model system.

The release of five cichlid genomes provides further opportunity for the molecular characterization of diversification. The five sequenced species encompass the phylogenetic and geographic diversity of East African cichlids (*18*). These genomes will serve as important resources, anchoring points, and templates for comparative genomic studies.

Sequencing of many more genomes, from many more species, will help to determine the contribution of mutation, selection, drift, and migration to diversification. This endeavor would also allow the detection of regulatory and coding polymorphisms that segregate in natural populations, which in turn would facilitate the linking of genotypes to phenotypes. East African cichlid fishes thus offer the possibility to dissect the interplay of thousands of genes from many genomes, found in many cells, forming tissues in many individuals, in many popula-

PHOTO CREDITS: M. LOMBARDOI AND N. *TRETOCEPHALUS*, WIKIMEDIA.ORG; L. *LABIATUS*, M. MUSCHICK/UNIV. OF SHEFFIELD; ALL OTHERS, E. SCHRAML

tions, encompassing hundreds of species that occupy various ecological niches across replicate adaptive radiations.

To keep up with these advances on the molecular and genomic aspects of cichlid diversification, it will be important to increase the efforts at the organismal and life-history level by surveying ecology, morphology, and behavior. This integration would make cichlids a role model not only for adaptive radiation and explosive speciation but also for the survey of interactions at all levels of biological organization.

References

1. Y. H. Loh, S. V. Yi, J. T. Streelman, *Genome Biol. Evol.* **3**, 55 (2011).
2. L. Baldo, M. E. Santos, W. Salzburger, *Genome Biol. Evol.* **3**, 443 (2011).
3. R. Miyagi *et al.*, *Mol. Biol. Evol.* **29**, 3281 (2012).
4. C. E. Wagner, L. J. Harmon, O. Seehausen, *Nature* **487**, 366 (2012).
5. M. Colombo *et al.*, *Mol. Ecol.* 10.1111/mec.12029 (2012).
6. M. R. Kidd, N. Duftner, S. Koblmüller, C. Sturmbauer, H. A. Hofmann, *PLoS ONE* **7**, e31236 (2012).
7. T. Manousaki *et al.*, *Mol. Ecol.* 10.1111/mec.12034 (2012).
8. D. Schluter, *The Ecology of Adaptive Radiation* (Oxford Univ. Press, New York, 2000).
9. S. Gavrilets, J. B. Losos, *Science* **323**, 732 (2009).
10. T. D. Kocher, *Nat. Rev. Genet.* **5**, 288 (2004).
11. G. J. Fraser *et al.*, *PLoS Biol.* **7**, e31 (2009).
12. R. B. Roberts *et al.*, *Science* **326**, 998 (2009).
13. R. C. Albertson, J. T. Streelman, T. D. Kocher, P. C. Yelick, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 16287 (2005).
14. K. Fujimura, T. D. Kocher, *Aquaculture* **319**, 342 (2011).
15. G. Fryer, T. D. Iles, *The Cichlid Fishes of the Great Lakes of Africa: Their Biology and Evolution* (Oliver & Boyd, Edinburgh, 1972).
16. G. W. Barlow, *The Cichlid Fishes. Nature's Grand Experiment in Evolution* (Perseus, Cambridge, MA, 2000).
17. O. Seehausen *et al.*, *Nature* **455**, 620 (2008).
18. Four members of radiating clades were sequenced, plus a sister taxon, the Nile tilapia. See www.broadinstitute. org/models/tilapia.

10.1126/science.1224818

PHYSICS

Quantum Procrastination

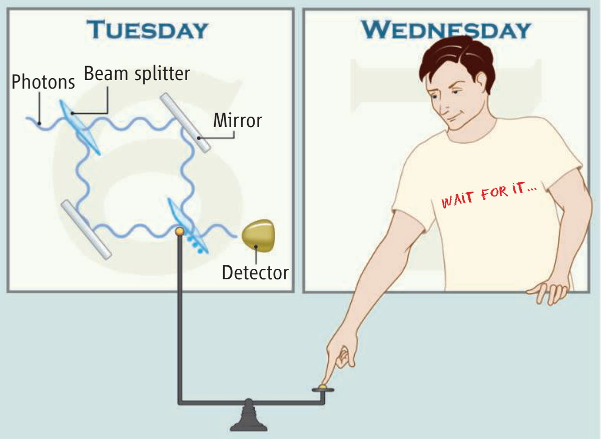
Entangling two photons allows the wave and particle nature of light to be interchanged even after the light has already been detected.

Seth Lloyd

Do you have a decision you have to make but you just can't bring yourself to do it? As the irrevocable moment approaches, you squirm more and more, but something inside you says, "Not now, not yet." Then when it's already almost too late, in a burst of energy and shame, you come through—or not. Afterward, you are irrationally resentful, as if someone other than yourself is responsible for disturbing your peace of mind. You vow that the next time a decision arises, you will make it expeditiously. If you are a severe procrastinator like me (at least when it came to starting this article), have hope—quantum mechanics is coming to your rescue. On pages 637 and 634 of this issue, experiments by Kaiser *et al*. (*1*) and Peruzzo *et al*. (*2*) show that in the presence of quantum entanglement (in which outcomes of measurements are tied together), it is possible to hold off making a decision, even if events seem to have already made one. Quantum procrastination ("proquastination") allows you to put off for tomorrow what you should have done today.

The experiments are based on Wheeler's famous delayed-choice experiment (*3*). Although photons are particles of light, they also possess a wavelike nature and can exhibit interference effects. Suppose that the path lengths of a Mach-Zehnder interferometer (*4*, *5*) have been tuned to make the photon come out of one port of the final beam splitter with probability 1 (see the figure). After the photon has passed the first beam splitter, so that it is fully inside the interferometer, and before it has reached the second beam splitter, you decide to whisk away that second beam splitter, preventing any interference between the photon's two paths from taking place. Without interference, the photon behaves like a particle and emerges with equal probability out of either of the two ports of the apparatus where the second beam splitter used to be.

If instead you choose to leave the beam splitter in, the wavelike nature of the photon asserts itself to exhibit interference between the two paths that the single particle takes in quantum superposition, and the photon would emerge from only one port with probability 1. That is, even though you have delayed the choice of removing the beam splitter until after the photon—if it really were a classical particle—should be traveling along one path or the other, by restoring the beam splitter, you can reinstate the photon's wavelike nature and have it report that it was traveling along both paths simultaneously.



Welcomed delays. Two studies use quantum entanglement in delayed choice experiments; the outcome for the first photon detected (whether it is a particle or a wave or has intermediate character) is determined by later measurements. Kaiser *et al*. entangle the first photon's polarization with that of the second photon, so that its outcome depends on the second photon's polarization. Peruzzo *et al*. entangle the photon with the presence or absence of a beam splitter in the setup and again delay the outcome of the first photon's state. If the photon states could be stored in quantum memories, it might be possible to delay the outcome of the first photon detection (on a Tuesday) until the observer makes a choice on Wednesday.

Since Wheeler proposed his delayed-choice gedanken experiment in 1984, a horde of theories and experiments exhibiting weird quantum effects has spread across the scientific landscape, including experimental demonstrations of Wheeler's proposal (*6*). Quantum information theory has supplied a general language for discussing such quantum weirdness, and small but effective quantum information processors have provided the wherewithal to demonstrate virtually any effect of quantum superposition and entanglement on a small number of quantum bits (*7*). As effects such as Wheeler's delayed-choice experiment and its relatives, such as the quantum eraser (*8*), have become commonplace, they have lost some of their power to amaze.

Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. E-mail: slloyd@mit.edu

Published by AAAS

Chapter 6

The ecological and genetic basis of convergent thick-lipped phenotypes in cichlid fishes

Marco Colombo, Eveline T. Diepeveen, Moritz Muschick, M. Emília Santos, Adrian Indermaur, Nicolas Boileau, Marta Barluenga and Walter Salzburger

Published in Molecular Ecology Special Issue “Ecological Genomics”

Personal contribution:
In this study I contributed to the *Lobochilotes labiatus* sampling, gene expression study design, gene expression data analysis, and manuscript preparation

MOLECULAR ECOLOGY

Molecular Ecology (2013) **22**, 670–684 doi: 10.1111/mec.12029

The ecological and genetic basis of convergent thick-lipped phenotypes in cichlid fishes

MARCO COLOMBO,*[1] EVELINE T. DIEPEVEEN,*[1] MORITZ MUSCHICK,*‡
M. EMILIA SANTOS,* ADRIAN INDERMAUR,* NICOLAS BOILEAU,* MARTA BARLUENGA† and WALTER SALZBURGER*

**Zoological Institute, University of Basel, Vesalgasse 1, 4051, Basel, Switzerland, †Museo Nacional de Ciencias Naturales, CSIC, José Gutierrez Abascal 2, 28006, Madrid, Spain*

Abstract

The evolution of convergent phenotypes is one of the most interesting outcomes of replicate adaptive radiations. Remarkable cases of convergence involve the thick-lipped phenotype found across cichlid species flocks in the East African Great Lakes. Unlike most other convergent forms in cichlids, which are restricted to East Africa, the thick-lipped phenotype also occurs elsewhere, for example in the Central American Midas Cichlid assemblage. Here, we use an ecological genomic approach to study the function, the evolution and the genetic basis of this phenotype in two independent cichlid adaptive radiations on two continents. We applied phylogenetic, demographic, geometric morphometric and stomach content analyses to an African (*Lobochilotes labiatus*) and a Central American (*Amphilophus labiatus*) thick-lipped species. We found that similar morphological adaptations occur in both thick-lipped species and that the 'fleshy' lips are associated with hard-shelled prey in the form of molluscs and invertebrates. We then used comparative Illumina RNA sequencing of thick vs. normal lip tissue in East African cichlids and identified a set of 141 candidate genes that appear to be involved in the morphogenesis of this trait. A more detailed analysis of six of these genes led to three strong candidates: *Actb*, *Cldn7* and *Copb*. The function of these genes can be linked to the loose connective tissue constituting the fleshy lips. Similar trends in gene expression between African and Central American thick-lipped species appear to indicate that an overlapping set of genes was independently recruited to build this particular phenotype in both lineages.

Keywords: adaptive radiation, cichlid species flocks, convergent evolution, East Africa, ecological genomics, RNAseq

Received 9 March 2012; revision received 4 July 2012; accepted 15 July 2012

Introduction

Adaptive radiation is the rapid evolution of an array of species from a common ancestor as a consequence of the emerging species' adaptations to distinct ecological niches (Simpson 1953; Schluter 2000; Gavrilets & Losos 2009). It is typically triggered by ecological opportunity in form of underutilized resources—just as being provided after the colonization of a new habitat, the extinction of antagonists and/or the evolution of a novel trait, which is then termed an evolutionary 'key innovation' (Gavrilets & Vose 2005; Gavrilets & Losos 2009; Losos & Ricklefs 2009; Losos 2010; Yoder *et al.* 2010; Matschiner *et al.* 2011). Whatever the circumstances were that initiated an adaptive radiation, there is always a strong link between adaptively relevant traits and the habitat and/or foraging niche (a 'phenotype–environment correlation'; Schluter 2000). In the most illustrative examples of adaptive radiation, the Darwin's finches on the Galapagos archipelago, the *Anolis* lizards on the

Correspondence: Walter Salzburger, Fax: +41 61 267 0301;
E-mail: walter.salzburger@unibas.ch

‡ Present address: Department of Animal and Plant Sciences, The University of Sheffield, Sheffield, S10 2TN, UK.

[1] These authors contributed equally to this work.

Caribbean islands and the cichlid fishes of the East African Great Lakes, this correlation exists between beak-shape and food source (finches), limb morphology and twig diameter (anoles), and the architecture of the mouth and jaw apparatus and foraging mode (cichlids) (Schluter 2000; Butler *et al.* 2007; Grant & Grant 2008; Losos 2009; Salzburger 2009).

An interesting aspect of many adaptive radiations is the frequent occurrence of convergent (or parallel) evolution (Schluter & Nagel 1995; Harmon *et al.* 2005; Arendt & Reznick 2008; Losos 2011; Wake *et al.* 2011). For example, similar ecotype morphs of anoles lizards have evolved independently on different Caribbean islands (Losos *et al.* 1998; Harmon *et al.* 2005; Losos & Ricklefs 2009), benthic–limnetic and lake–stream species pairs of threespine sticklebacks emerged repeatedly in and around postglacial lakes (Rundle *et al.* 2000; Berner *et al.* 2010; Roesti *et al.* 2012), and a whole array of convergent forms of cichlid fish emerged between the lakes of East Africa (Kocher *et al.* 1993; Salzburger 2009). Such instances of convergent evolution are generally interpreted as the result of the action of similar selection regimes in isolated settings (Schluter & Nagel 1995; Rundle *et al.* 2000; Nosil *et al.* 2002; Harmon *et al.* 2005; Losos 2011). It has further been suggested that if radiations are truly replicated (i.e. driven by adaptive processes), convergence in morphology should tightly be associated with convergence in ecology and behaviour (Johnson *et al.* 2009).

The species flocks of cichlid fishes in the East African Great Lakes Victoria, Malawi and Tanganyika represent the most species-rich extant adaptive radiations in vertebrates (Kocher 2004; Seehausen 2006; Salzburger 2009). Several hundreds of endemic cichlid species have emerged in each lake within a period of several millions of years (as is the case for Lake Tanganyika; Salzburger *et al.* 2002; Genner *et al.* 2007) to <150 000 years (as in Lake Victoria; Verheyen *et al.* 2003). The various endemic cichlid species differ greatly in the morphology of the trophic apparatus (mouth form and shape, jaw structure and dentition) as well as in coloration and pigmentation, suggesting that both natural and sexual selection are jointly responsible for adaptive radiation and explosive speciation in cichlids (Salzburger 2009). Interestingly, convergent forms that emerged in independent cichlid adaptive radiations often show very similar coloration patterns in addition to matching body shapes and mouth morphologies (Kocher *et al.* 1993; Stiassny & Meyer 1999; Salzburger 2009). This has led to speculations whether selection alone is sufficient to explain convergence, or whether genetic or developmental constraints have contributed to the morphogenesis of these matching phenotypes (Brakefield 2006).

The present study focuses on the morphology, ecology and the genetic basis of a peculiar mouth trait in cichlid fishes, which has evolved multiple times: hypertrophied ('fleshy') lips (see Box 1 in Salzburger 2009). The exact function of the thick lips in cichlids is unknown, although this feature is generally implicated in a specific foraging mode (Fryer 1959; Fryer & Iles 1972; Arnegard *et al.* 2001). Fleshy lips are often interpreted as an adaptation for feeding on invertebrates and crustaceans hidden in crannies, with the lips being used to seal cracks and grooves to facilitate the sucking of prey (Barlow & Munsey 1976; Ribbink *et al.* 1983; Seehausen 1996; Konings 1998). Alternatively, it has been suggested that hypertrophied lips protect from mechanical shocks (Greenwood 1974; Yamaoka 1997), and that they function as taste receptors (Arnegard *et al.* 2001) or as mechanoreceptors (Fryer 1959; Fryer & Iles 1972). [Note, however, that there is no increase in sensory cells in lip tissue (Greenwood 1974).]

It is remarkable that thick-lipped species appear to be a common outcome of cichlid adaptive radiations. For example, the large cichlid assemblages in East Africa all contain at least one such taxon (Lake Victoria: *Haplochromis chilotes*; Lake Malawi: *Chilotilapia euchilus, Abactochromis labrosus, Otopharynx pachycheilus, Placidochromis milomo, Protomelas ornatus*; Lake Tanganyika: *Lobochilotes labiatus*). In addition, cichlids featuring hypertrophied lips are known from, for example, the Midas Cichlid (*Amphilophus* spp.) assemblage in the large lakes of Nicaragua, where a thick-lipped species (*A. labiatus*) is common in rocky habitats (Fig. 1). Occasionally, hypertrophied lips are also observed in other related cichlids in Nicaragua, such as in the riverine species *Tomacichla tuba* (Villa 1982) or in *Astatheros rostratus* (pers. obs.). Additional riverine representatives with hypertrophied lips are also found in South America (*Crenicichla tendybaguassu*) and Western Africa (*Thoracochromis albolabris*). Hypertrophied lips are not unique to cichlids, though. For example, the adaptive radiation of the sailfin silverside fish (Telmatherinidae) in the Malili lakes of Sulawesi (Herder *et al.* 2006) and the barbs of Lake Tana in Ethiopia (Sibbing *et al.* 1998; de Graaf *et al.* 2008) also produced thick-lipped species.

Members of the family Cichlidae are distributed in the Southern hemisphere, with a few ancestral lineages in India, Sri Lanka and Madagascar and two exceptionally species-rich clades, one in Central and South America and one in Africa (Salzburger & Meyer 2004). This biogeographical pattern is consistent with a Gondwanan origin of the Cichlidae, dating the split between American and African representatives to ~100 Ma (Salzburger & Meyer 2004; Sereno *et al.* 2004; Genner *et al.* 2007). This set-up opens the possibility to study the ecological and genetic basis of a convergent trait across one of the

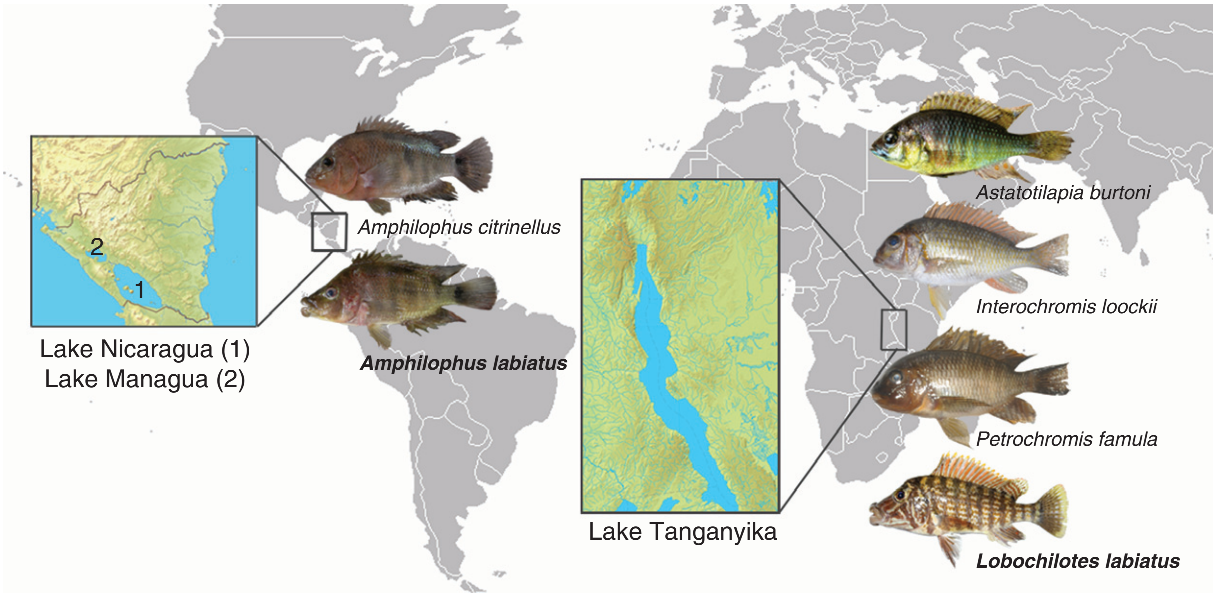


Fig. 1 Map of the Southern hemisphere showing the two study systems, the Midas Cichlid (*Amphilophus* sp.) species complex in Nicaragua, Central America, and the Tropheini in Lake Tanganyika, East Africa.

largest possible phylogenetic and geographical distances in cichlids and, hence, in the complete absence of gene flow and outside the influence of ancestral polymorphism and/or standing genetic variation.

Here, we applied an integrative approach in two cichlid fish radiations, the one of the Tropheini in East African Lake Tanganyika and the Midas Cichlid assemblage in Nicaragua, to uncover the ecological and genetic basis of the thick-lipped phenotype. More specifically, we compared the two '*labiatus*' species to one another and to their sister species by means of geometric morphometric and stomach content analyses; we placed them in their respective radiations by phylogenetic and demographic analyses; and we provide field observations on foraging strategies for one of them (*L. labiatus*). To study the genetic basis of hypertrophied lips, we first applied comparative transcriptome analyses (RNA-seq) on the basis of Illumina next-generation sequencing of juvenile and adult individuals of the African species *L. labiatus* (in comparison with a closely related species for which a genome sequence is available). In a second step, we tested candidate genes identified by RNAseq in representatives of both radiations in a quantitative real-time PCR environment.

Materials and methods

Study species

This study focuses on two thick-lipped species, *Lobochilotes labiatus* from East African Lake Tanganyika and *Amphilophus labiatus* from Nicaragua. *Lobochilotes labiatus* is a member of the rock-dwelling Tanganyikan cichlid tribe Tropheini and therefore part of the most species-rich group of cichlids, the haplochromines, which include the Tanganyikan Tropheini, many riverine species and the species flocks of Lakes Victoria and Malawi (Salzburger *et al.* 2002, 2005). The Tropheini themselves underwent a subradiation within Lake Tanganyika (see e.g. Sturmbauer *et al.* 2003). *Amphilophus labiatus* is part of the Midas Cichlid assemblage in Nicaragua and occurs in the large Central American lakes Managua and Nicaragua, where it co-occurs with the most common species in the area, *A. citrinellus* (Barlow 1976; Barluenga & Meyer 2010). For this study, we sampled a total of 84 and 74 specimens of the Central American species *Amphilophus citrinellus* and *A. labiatus*, respectively, and 143 specimens of *L. labiatus* plus 14 additional Haplochromini/Tropheini specimens from Lake Tanganyika. Exact sampling locations and dates for specimens used for the genetic analysis and GenBank accession numbers are provided in Appendix S1.

Sampling, DNA and RNA extraction

Sampling of *L. labiatus* and other Tropheini species was performed between 2007 and 2011 in the Southern part of Lake Tanganyika, East Africa; *A. labiatus* and its congeners were collected in September 2009 in the two large Nicaraguan lakes Managua and Nicaragua (see Appendix S1 for details). Fishes were processed in the field following our standard operating procedure: fishes were individually labelled, measured (total and standard length) and weighted and a photograph was taken from the left side of each specimen using a Nikon P5000 or a Nikon D5000 digital camera (fins were spread out using clips); then, a piece of muscle tissue and a fin-clip were taken as DNA sample and preserved in ethanol; fishes were then dissected and RNA samples from lip and other tissues were preserved in RNAlater (Ambion); the whole intestinal tract was removed and stored in ethanol.

For DNA extraction, we either applied a high-salt extraction method (Bruford *et al.* 1998) or used a MagnaPure extraction robot (Roche, Switzerland) following the manufacturer's protocol. RNA was extracted according to the Trizol method with either Trizol (Invitrogen) or TRI reagent (Sigma). Lip tissue was homogenized with a PRO200 Homogenizer (PRO Scientific Inc.) or with a BeadBeater (FastPrep-24; MP Biomedicals). DNase treatment following the DNA Free protocol (Ambion) was performed to remove any genomic DNA from the samples. Subsequent reverse transcription was achieved by using the High Capacity RNA-to-cDNA kit (Applied Biosystems). For the *A. burtoni* samples, up to two individuals (adults) or up to eight individuals (juveniles) were used per sample, due to a diminutive amount of lip tissue extracted from these fishes. All other samples were taken from a single specimen.

Phylogenetic and demographic analyses

We first wanted to phylogenetically place the thick-lipped species into the respective clade of East African and Nicaraguan cichlids. We thus performed a phylogenetic analysis of the Tanganyikan cichlid tribe Tropheini (see also Sturmbauer *et al.* 2003) and used haplotype genealogies to reconstruct the evolutionary history in the much younger *Amphilophus* species assemblage in Nicaragua, where phylogenetic analyses are not expedient due to the lack of phylogenetic signal (see also Barluenga *et al.* 2006; Barluenga & Meyer 2010). We also performed mismatch analyses within *A. citrinellus*, *A. labiatus* and *L. labiatus* to compare their demographic histories.

We amplified three gene segments for each of the three focal species and additional Tropheini/Haplochromini species: the first segment of the noncoding mtDNA control region and two nuclear loci containing coding and noncoding DNA (a segment each of the *endothelin receptor 1*, *ednrb1* and the *phosphatidin phosphatase 1*, *phpt1*). We used previously published primers L-Pro-F (Meyer *et al.* 1994) and TDK-D (Lee *et al.* 1995) for the control region and ednrb1F and ednrb1R (Lang *et al.* 2006) for *ednrb1*, and so far unpublished primers 38a_F (5′-AGC AGG GTT GAC CTT CTC AA-3′) and 38a_R (5′-TGG CTA AAA TCC CCG ATG TA-3′) for *phpt1*. Polymerase chain reaction (PCR) amplification, purification and cycle sequencing were performed as described elsewhere (Diepeveen & Salzburger 2011); an ABI 3130*xl* capillary genetic analyzer (Applied Biosystems) was used for DNA sequencing.

The resulting sequences were complemented with already available sequences. In the case of the Tropheini, we also included available sequences of the mitochondrial NADH dehydrogenase subunit 2 gene (ND2) (see Appendix S1 for GenBank accession numbers). Sequences were aligned with MAFFT (Katoh & Toh 2008) resulting in a total length of 2345 bp for the Tropheini (control region: 371 bp; ND2: 1047 bp; ednrb1: 538 bp; *phpt1:* 389 bp) and 1620 bp for *Amphilophus* (control region: 371 bp; ednrb1: 743 bp; *phpt1:* 469 bp). Maximum-likelihood and Bayesian inference phylogenetic analyses of the Tropheini were performed for each gene segment separately (not shown) and for a concatenated alignment with PAUP* (Swofford 2003) and MRBAYES (Ronquist & Huelsenbeck 2003), respectively. The appropriate model of sequence evolution was detected with JMODELTEST (Posada 2008) applying the Akaike Information Criterion (AIC). A maximum-likelihood bootstrap analysis with 100 pseudoreplicates was performed in PAUP*, and MR. BAYES was run for eight million generations with a sample frequency of 100 and a burn-in of 10%. We then used MESQUITE (www.mesquiteproject.org) to map feeding specializations on the resulting maximum-likelihood topology and to reconstruct ancestral character states with parsimony. Data on feeding mode from the Haplochromini/Tropheini species other than *L. labiatus* are based on Brichard (1989), Nori (1997), Yamaoka (1997) and Konings (1998).

Haplotype genealogies for the *Amphilophus* data set were constructed following the method described in the study by Salzburger *et al.* (2011) on the basis of a maximum-likelihood tree and sequences of the mitochondrial control region and the nuclear *ednrb1* gene (*phpt1* was not used here due to the limited number of haplotypes found). Mismatch analyses were performed on the basis of mtDNA sequences with ARLEQUIN 3.0 (Excoffier *et al.* 2005).

Geometric morphometric analyses

In order to test for similarities in overall body shape between the thick-lipped forms from Central America and East Africa, we performed geometric morphometric analyses on the basis of digital images. Body shape was quantified in a set of 58 *A. citrinellus*, 27 *A. labiatus* and 27 *L. labiatus* using 17 homologous landmarks (see Appendix S2; note that lip shape was not assessed to prevent a bias). Data acquisition was carried out using TPSDIG (Rohlf 2006), and data were analysed with MORPHOJ (Klingenberg 2011). For all shape comparisons, we used the residuals of a within-species regression of shape on centroid size to reduce allometric effects within species, in order to retain shape differences between differently sized species. For the same reason, we only included *L. labiatus* individuals with a body size larger than 12 cm total length. We then performed a discriminant function analysis between all pairs of species and a principal component analysis (PCA). To identify morphological changes associated with the enlarged lip phenotype, we compared *A. labiatus* to its closest relative, *A. citrinellus*. In the case of *L. labiatus*, we made use of our new phylogeny of the Tropheini (Fig. 2a) and body shape data of *L. labiatus* and its nine closest relatives [*Petrochromis macrognathus*, *P. polyodon*, *P. ephippium*, *Lobochilotes labiatus*, *Simochromis diagramma*, *S. babaulti*, *Gnathochromis pfefferi*, *Pseudosimochromis curvifrons*, *Limnotilapia dardenni* and *Ctenochromis horei* (M. Muschick, A. Indermaur & W. Salzburger, unpublished data)] to reconstruct the landmark configuration of the direct ancestor to *L. labiatus*. This was carried out in MorphoJ using branch length-weighted squared-change parsimony. The changes in landmark configurations along a discriminant function (Nicaraguan species) or along the shape-change vector from the estimated ancestral shape to *L. labiatus* were increased threefold to produce Fig. 3. The shape differences between species shown in Fig. 3 accurately reflect the shape-change vectors for landmark positions. Outlines were interpolated and added to Fig. 3 to help the reader envision these shape differences in the context of fish body shape.

Stomach and gut content analyses

To assess trophic specialization of the thick-lipped cichlid species, we performed comparative stomach and gut content analyses. To this end, stomachs and guts were opened step-by-step. First, the stomach was opened and emptied under a binocular followed by the remaining parts of the intestine. All items were grouped into seven food categories: hard-shelled (crustaceans, snails, mussels), small arthropods (insects and zooplankton), fish scales, fish remains, plant seeds and plant material other than seeds. For each specimen, the wet weight of each food category was measured on a Kern ALS 120-4 scale (Kern, Germany) and was then used to calculate Schoener's index of proportional diet overlap (Schoener 1970). We analysed stomach and gut contents in a total of 159 specimens: *A. citrinellus* ($N = 58$; of which 25 had contents), *A. labiatus* ($N = 62$; 34) and *L. labiatus* ($N = 39$; 29). We note that such an analysis has the drawback that it only covers food uptake in the last few hours or days before sampling.

Field observations in Lobochilotes labiatus

The feeding behaviour of *L. labiatus* was observed at our field site near Mpulungu, Zambia, in concrete ponds (1.5 × 1.5 × 1 m). The purpose of these observations under semi-natural conditions and with wild specimens was to document if and how the lips are used in processing the main prey item identified in the stomach content analyses. The ponds were equipped with stones of ~20–30 cm diameters that covered the ground and formed caves as they occur naturally in the habitat of *L. labiatus*. Each pond was stocked with five to six freshly caught and unharmed adult individuals of *L. labiatus*. After an acclimatization period of at least 4 days, fish were offered snails of different sizes and their feeding behaviour was recorded with two underwater cameras (Canon Ixus 65 with WP-DC3 underwater case; Olympus μ tough-6000) for a period of 1 h each.

Comparative gene expression assays using RNAseq

For the identification of differentially expressed genes in thick-lipped species, we performed RNA sequencing (RNAseq) comparing lip tissue from a thick-lipped species to lip tissue from a reference species. We decided to perform these experiments in the African species *L. labiatus* and to use the closely related species *Astatotilapia burtoni* as reference taxon for several reasons such as the availability of laboratory strains and of sufficient RNA samples from adult and juvenile individuals. Most importantly, we chose this set-up because of the availability of various genomic resources for *A. burtoni*, such as a whole-genome sequence and a set of ~50 000 partly annotated expressed sequence tags (ESTs) (Salzburger *et al.* 2008; Baldo *et al.* 2011), which is crucial for the analysis and interpretation for RNAseq data. Such resources are currently not publicly available for *Amphilophus*.

In a first step, RNA was extracted from adult and juvenile individuals of *L. labiatus* and *A. burtoni* (see above for the RNA extraction protocol). RNA quality and quantity were determined on a NanoDrop 1000 spectrophotometer (Thermo Scientific) and by gel electrophoresis. RNA samples were pooled to create four samples subjected to RNA sequencing (RNAseq): (i) *A. burtoni* adult ($N = 3$); (ii) *A. burtoni* juvenile ($N = 1$); (iii) *L. labiatus* adult ($N = 2$); and (iv) *L. labiatus* juvenile ($N = 3$). Five micrograms of RNA per RNAseq sample was sent for Illumina sequencing at the Department of Biosystems Science and Engineering (D-BSSE), University of Basel and ETH Zurich. For library construction and sequencing, standard protocols were applied. Poly-A mRNA was selected using poly-T oligo-attached magnetic beads. The recovered mRNA was fragmented into smaller pieces using divalent cations under increased temperature. cDNA was produced using reverse transcriptase and random primers, followed by second-strand cDNA synthesis using DNA polymerase

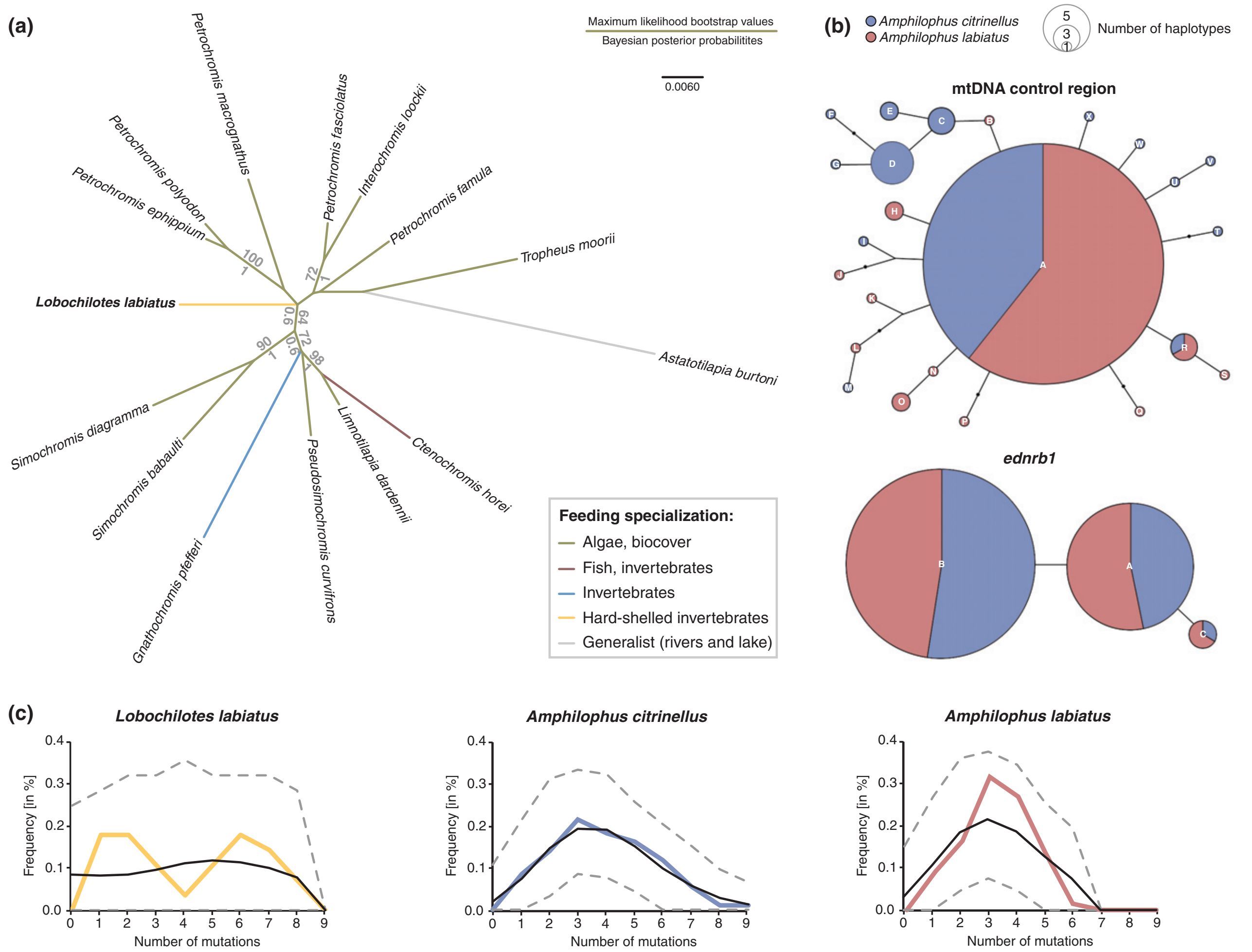


Fig. 2 Evolutionary origin of the thick-lipped species in East African Lake Tanganyika and in the Great Lakes of Nicaragua. (a) Maximum-likelihood tree of the Tropheini from Lake Tanganyika based on two mitochondrial (control region and ND2) and two nuclear (*ednrb1* and *phpt1*) gene segments (2345 bp in total) and the GTR+G+I model of molecular evolution. Numbers above the branches refer to maximum-likelihood bootstrap values, and numbers below are Bayesian posterior probabilities (note that support values are only shown for branches with bootstrap values >60). Branches are colour-coded according to feeding specializations; the trait values for internal branches have been reconstructed with MESQUITE. (b) Haplotype genealogies of the two *Amphilophus* species based on the mitochondrial control region and the nuclear *endrb1* gene. A large fraction of the haplotypes is shared between *A. citrinellus* and *A. labiatus*. (c) Results from the mismatch analysis on the basis of the mitochondrial control region showing the inferred demographic histories for *L. labiatus*, *A. citrinellus* and *A. labiatus*. Coloured lines represent the observed data, the black line indicates the best-fit model, and the dashed lines in grey indicate the upper and lower boundaries from the simulations in ARLEQUIN.

I and RNaseH. cDNA went through an end-repair process, the addition of a single 'A' base and ligation of the adapters. It was then purified and enriched with PCR to create the final cDNA library. Each library was sequenced in one lane on an Illumina Genome Analyzer IIx (read length was 76 bp). Illumina reads are available from the Sequence Read Archive (SRA) at NCBI under the accession number SRA052992.

The Illumina reads were assembled into three different data sets for further analyses: (i) a quality-filtered data set (Data set 1), where the quality of the reads was assessed with the FASTX toolkit tools implemented in GALAXY [version September/October 2011; available at http://main.g2.bx.psu.edu/ (Giardine *et al.* 2005; Blankenberg *et al.* 2010; Goecks *et al.* 2010)]; low-quality reads were discarded applying quality filter cut-off values of 22–33. (ii) a quality-filtered plus trimmed data set (Data set 2), in which all the reads were trimmed to a length of 42 bp to evaluate the effects of read length (iii) as a control for the effect of trimming and filtering, a nonquality-filtered, nontrimmed data set (Data set 3).

The reads of the three data sets were then aligned to a reference cichlid assembly (Baldo *et al.* 2011) with NOVOALIGN 2.07.06 (http://www.novocraft.com/) after indexing the reference sequences with NOVOINDEX (http://www.novocraft.com/) using default parame-

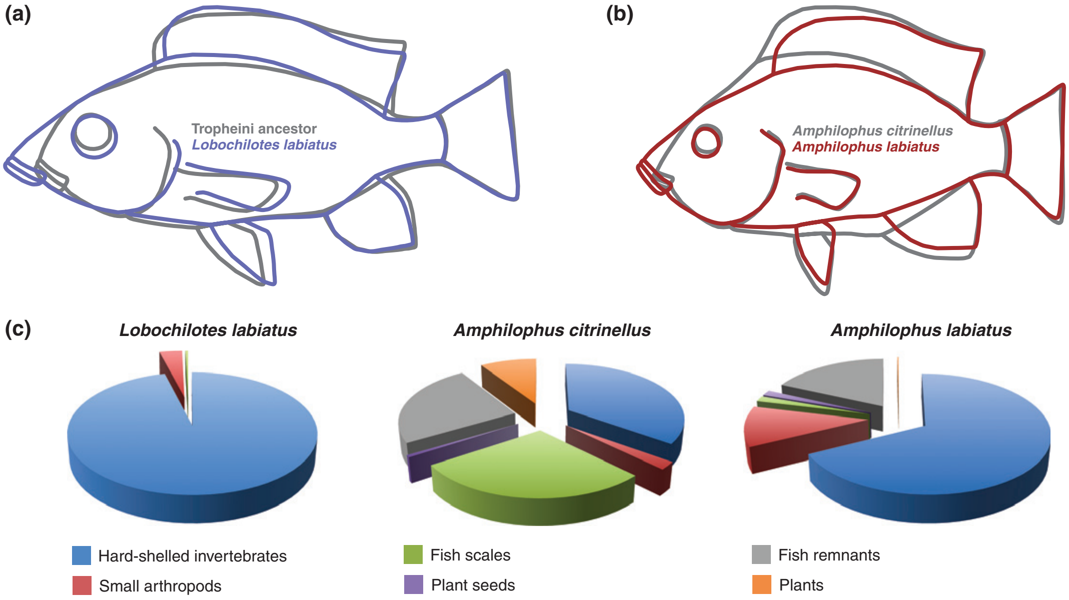


Fig. 3 Ecomorphology of the thick-lipped cichlid species in Central America and in Lake Tanganyika. (a) Body shape of *L. labiatus* in comparison with a reconstruction of the ancestor of *L. labiatus* and nine closely related Tropheini species. (b) Differences in body shape between *A. citrinellus* and *A. labiatus* along a discriminate function. In both plots, changes in landmark positions were increased threefold and interpolated outlines added for illustration purposes. Landmark locations are indicated in black on the reconstructed outlines in plot (a). (c) Analysis of stomach and gut content in the focal species. The fraction of each food category is shown.

ters. The alignment was performed using default settings with a maximum alignment score (t) of 180 and a maximum number of alignments for a single read (e) of 100; reads with multiple alignment locations were discarded. Next SAMtools version 0.1.18 (Li *et al.* 2009) was used to sort and index the files and to generate count files, which were subsequently transformed into count tables and analysed in the R package DESeq version 1.0.5 (Anders & Huber 2010). Differentially expressed genes between the four experimental groups were detected using a model based on a negative binomial distribution implemented in DESeq. Differentially expressed genes with *P*-values (adjusted for multiple testing) >0.05 and/or a quotient of variance >1.00 were discarded to reduce the number of false positives. The remaining differentially expressed genes of all pairwise comparisons were tested for multiple hits. Next the hits of the three data sets were compared with each other to create a candidate gene list, consisting of genes that were found in multiple analyses in all three data sets. Lastly, these hits were compared to the annotated *A. burtoni* ESTs of Baldo *et al.* (2011).

Comparative gene expression assays using quantitative real-time PCR

Based on their function according to gene ontology terms (GO terms; http://www.geneontology.org/) and their putative involvement in lip formation and/or hypertrophy in other organisms, six candidate genes were selected out of the list of differentially expressed genes for further characterization by means of quantitative real-time PCR (qPCR). These candidate genes are the *Bcl2 adenovirus e1b 19-kda protein-interacting protein 3* (*BNIP3*), *long-chain-fatty-acid(CoA)-ligase 4* (*ACSL4*), *histone 3.3* (*His3*), *beta actin* (*Actb*), *coatomer subunit beta* (*Copb*) and *claudin 7* (*Cldn7*; see Table 1 for primer details). qPCR experiments were performed in total of 36 cichlid specimens: *L. labiatus* (six adults, six juveniles), *A. burtoni* (six adults, six juveniles), *A. labiatus* (six adults) and *A. citrinellus* (six adults). By performing two pairwise comparisons between a thick-lipped and a normal-lipped species (a species pair each from Africa and Nicaragua), we effectively control for species-specific expression differences, as genes specific to thick-lip tissue should be upregulated in both comparisons.

The experiments were conducted on a StepOnePlus Real-Time PCR system (Applied Biosystems) as described elsewhere (Diepeveen & Salzburger 2011) using the *elongation factor 1* (*EF1*) and the *ribosomal protein SA3* (*RpSA3*) as endogenous controls. Average relative quantifications (RQ) were calculated for the six experimental groups and subsequently analysed with a two-tailed unpaired t-test using GraphPad Prism version 5.0a for Mac OS X (www.graphpad.com). We compared the expression levels between the two thick-lipped species and a closely related normally lipped species (i.e. *L. labiatus* vs. *A. burtoni* and *A. labiatus* vs. *A. citrinellus*). We also compared adults vs.

Table 1 Primers used for the quantitative real-time PCR experiments

Locus	Forward (5′–3′)	Reverse (5′–3′)
Actb	CAGGCATCAGGGTGTAATGGTT	CAGGCATCAGGGTGTAATGGTT
Copb	GAGGCTACCTTGGCTGTCAAAG	GTGCTGGATGGTTTGAGGGTAA
His3	CATCTACTGGTGGAGTGAAGAAACC	GGATCTCACGCAGAGCAACA
ACSL4	TGGTTCTGCACCGGAGATG	TCTTGCGGTCAACAATTTGTAGA
BNIP3	AACAGTCCACCAAAGGAGTTCCT	CCTGATGCTGAGAGAGGTTGTG
Cldn7	GACATCATCCGGGCCTTCT	CACCGAACTCATACTTAGTGTTGACA
EF1	GCCCCTGCAGGACGTCTA	CGGCCGACGGGTACAGT
RpSA3	AGACCAATGACCTGAAGGAAGTG	TCTCGATGTCCTTGCCAACA

juveniles in the African species, as hypertrophy in lips is much less pronounced at juvenile stages, so that this experiment also captures ontogenetic changes in lip formation. As primer efficiency was lower in the Nicaraguan samples, no direct comparisons between African and Nicaraguan tissues were possible.

Results

Phylogenetic and demographic analyses

Our phylogenetic analysis of members of the Tanganyikan cichlid tribe Tropheini based on two mitochondrial and two nuclear DNA gene segments reveals only limited phylogenetic resolution between the main lineages of the tribe (Fig. 2a). This confirms an earlier analysis based on mitochondrial DNA only, which attributed the star-like phylogeny of the Tropheini to the rapidity of lineage formation in the early stages of the adaptive radiation of this clade (Sturmbauer *et al.* 2003). Just as in the previous study, the thick-lipped species *L. labiatus* represents a separate lineage (without a closely related sister-taxon) that branches off relatively early in the phylogeny, but shows affinities to the algae-eating genera *Petrochromis* and *Simochromis*.

The haplotype genealogies of the *Amphilophus* samples based on the mitochondrial control region and the nuclear *ednrb1* gene (Fig. 2b) revealed haplotype sharing between *A. citrinellus* and *A. labiatus* (see also Barluenga & Meyer 2010). While all *Amphilophus* sequences were identical in *phpt1*, we detected three shared haplotypes in ednrb1 and 24 haplotypes in the mitochondrial control region (two shared, ten unique to *A. labiatus* and twelve unique to *A. citrinellus*).

The mismatch analyses based on the mitochondrial control region sequences revealed unimodal distributions for the two sympatrically occurring *Amphilophus* species and a bimodal distribution for *L. labiatus* (Fig. 2c). According to this analysis, the demographic expansion of the two *Amphilophus* species happened at similar times, with the one of *A. citrinellus* being slightly older than that of *A. labiatus* (mean number of differences: 3.9 vs. 3.2; τ: 3.9 vs. 3.5; see also Barluenga & Meyer 2010, who provide a relative time frame for the evolution of the Midas Cichlid species complex); the mean number of differences in *L. labiatus* was 6.4 (τ: 6.5).

Geometric morphometric analyses

The PCA of overall body shape revealed substantial overlap between the two Nicaraguan species *A. citrinellus* and *A. labiatus* (Appendix S3). The African thick-lipped species *L. labiatus* is separated from these mainly by principal component 1 (accounting for 20.2% of the variance), whereas principal component 2 (covering 16.0% of the variance) did not discriminate much between species. The discriminant function analysis, in which we compared species in a pairwise manner, revealed the main morphological differences between species. Of the two Nicaraguan species, *A. labiatus* had a more acute head, less deep body and a larger mouth than *A. citrinellus* (Fig. 3) (see also Klingenberg *et al.* 2003). These characters were even more pronounced in *L. labiatus*, when compared to either of the *Amphilophus* species. However, the distance in morphospace between the two species with fleshy lips was somewhat smaller than between *A. citrinellus* and *L. labiatus* (procrustes distance 0.08 and 0.1, respectively). We also estimated the body shape of the ancestor of *L. labiatus* and the 9 most closely related Tropheini species. A comparison of this reconstructed shape and the mean shape of our *L. labiatus* samples highlighted similar morphological differences as the comparison of the Nicaraguan species (Fig. 3), especially in the mouth region.

Stomach and gut content analyses

The fractions of food categories in guts and stomachs differed between *A. citrinellus*, *A. labiatus* and *L. labiatus* (Fig. 3c). While the diet of *A. citrinellus* did not overlap with that of *A. labiatus* (Schoener's index: 0.58) or *L. labiatus* (Schoener's index: 0.38), we found significant overlap between the two thick-lipped species *A. labiatus* and *L. labiatus* (Schoener's index: 0.71) (note that any value >0.6 is considered 'biologically significant'; see Wallace 1981). The stomach and gut contents of both thick-lipped species consisted of a substantial fraction of hard-shelled prey (*Lobochilotes labiatus* 96%, *Amphilophus labiatus* 67.6%, *Amphilophus citrinellus 35%*).

Field observations in Lobochilotes labiatus

A careful inspection of the video material confirmed the findings from the stomach and gut content analyses that *L. labiatus* regularly feeds on snails (more than 90% of the stomach and gut content of *L. labiatus* consisted of snail shells). Small snails were engulfed using suction feeding without the lips touching the prey item or the surface (rocks) on which the items were placed. When feeding on larger snails, however, *L. labiatus* exhibited a different feeding strategy and snails were no longer taken up using suction feeding. Instead, *L. labiatus* used their lips to snatch the snails and they turned the snails a few times before they either swallowed the snails or spat them out (see Appendix S4).

Comparative gene expression assays using RNAseq

On average, ca. 42 million total reads were retrieved for each of the four RNAseq samples (*A. burtoni* adult, *A. burtoni* juvenile, *L. labiatus* adult and *L. labiatus* juvenile). Quality filtering and trimming reduced this number so that on average 21.9 (Data set 1), 24.6 (Data set 2) and 23.5 (Data set 3) million reads were aligned to the reference cichlid assembly. Five different pairwise comparisons were made to obtain genes that are differentially expressed between thick lips and normal lips (see Table 2 for the three comparisons with the highest number of genes being different). The largest number of differentially expressed genes between *L. labiatus* and *A. burtoni* was detected in adult lip tissue, with the majority of the genes being upregulated in *L. labiatus*. The total number of differentially expressed genes ranged from 9050 (Data set 3; three pairwise comparisons) to 15230 (Data set 2; five pairwise comparisons). A substantial fraction of these differentially expressed genes appeared in at least two comparisons in each data set (Data set 1: 2085 [22.1% of all hits]; Data set 2: 8078 [53.0%]; Data set 3: 1693 [18.7%]). Of these 'multiple hits', 1463 were detected in all three data sets and 560 of those could be unequivocally annotated.

A more stringent analysis, in which only loci that appeared in at least three of five comparisons were included, resulted in 231 differentially expressed genes. A functional annotation of these 231 hits with Blast2GO resulted in a total of 141 annotations (122 upregulated and 19 downregulated in *L. labiatus*; see Appendix S3). Based on their annotations, known functions and/or exceptional fold change (>1000) between *A. burtoni* and *L. labiatus*, thirteen genes were identified as good candidates for being involved in the morphogenesis of fleshy lips (Table 3).

Comparative gene expression assays using quantitative real-time PCR

The results of the comparative gene expression assays between the thick-lipped species and the normal-lipped species are depicted in Fig. 4 and Appendix S5. Overall, the qPCR experiments largely validate differential gene expression in normal and hypertrophied lip tissue as indicated by RNAseq. In the African species pair *L. labiatus* and *A. burtoni*, which were the two species used for RNAseq, differences were highly significant in four of the six genes tested: *Actb* ($P = 0.0099$), *Cldn7* ($P = 0.004$), *ACSL4* ($P = 0.0005$) and *His3* ($P = 0.0003$). However, we would like to point out one inconsistency between RNAseq and qPCR. *Actb* was actually found to be downregulated in hypertrophied lips by RNAseq, while it shows significantly higher expression levels in lip tissue in the qPCR experiments (Fig. 4).

The comparison between lip tissue in adult and juvenile *L. labiatus* and *A. burtoni* further revealed a trend towards higher expression in lip tissue of adult *L. labiatus* in *Actb*, *BNIP3*, *Cldn7* and *Copb* (Appendix S5), whereas, generally, an opposite trend is observed in *A. burtoni*, although statistical support was only found in two cases [*Cldn7* ($P = 0.0063$) and *ACSL4* ($P = 0.0328$)]. This again suggests that these genes are involved in the formation of fleshy lips. In the Nicaraguan species pair, a similar trend was observed as in the African species pair, with four of the five genes tested appearing to be upregulated in lip tissue

Table 2 Pairwise comparisons of differentially expressed genes and total number of unique differentially expressed genes in the three data sets compiled in this study

Comparison	Data set 1	Data set 2	Data set 3
AB vs. LL	7120 (4606; 2514)	7080 (4689; 2391)	7285 (4665; 2620)
AB vs. LLjuv	3611 (3395; 216)	13747 (10683; 3064)	2618 (2514; 104)
ABjuv vs. LLjuv	1116 (792; 324)	3971 (2710; 1261)	986 (687; 298)
Total	9407	15225	9050

AB, *Astatotilapia burtoni;* LL, *Lobochilotes labiatus*; juv, juvenile; numbers in brackets denote the number of upregulated and downregulated genes in *L. labiatus*.

Table 3 Thirteen candidate loci for the genetic basis of lip development in the East African cichlid *Lobochilotes labiatus*, based on RNAseq and qPCR in comparison with *Astatotilapia burtoni*, in combination with information on gene functions (in alphabetical order)

Locus	Abbreviation
ATPase mitochondrial precursor	*ATPmp*
Bcl2 adenovirus e1b 19-kda protein-interacting protein 3	*BNIP3*
Beta actin	*Actb*
Caspase-8	*Casp8*
Claudin 7	*Cldn7*
Coatomer subunit beta	*Copb*
Grainyhead-like protein 1 homolog	*Grhl1*
Heat-shock 70-kda protein 12a-like	*Hspa12al*
Histone 3.3	*His3*
Laminin subunit gamma-2	*Lamc2*
Long-chain-fatty-acid(CoA)-ligase 4	*ACSL4*
Sodium-dependent phosphate transporter 1	*Slc17a1*
Transcription factor ap-2 gamma	*Tfap2*

of *A. labiatus* as compared to *A. citrinellus* (Fig. 4; we could not amplify *BNIP3* here). We would like to note, however, that qPCR efficiency was less good in the *Amphilophus* samples, most likely because we used primers designed for the African species pair based on the available genomic resources, which also explains the limited statistical support for these comparisons. Interestingly, it seems that several loci (i.e. *Actb*, *Cldn7*, *Copb*, *His3*) are upregulated in both thick-lipped species when compared to their normally lipped relatives.

Discussion

The species flocks of cichlid fishes in the East African Great Lakes Victoria, Malawi and Tanganyika, counting hundreds of endemic species each, are prime examples of adaptive radiation and explosive speciation (see e.g. Kocher 2004; Seehausen 2006; Salzburger 2009). Interestingly, the cichlid adaptive radiations in East Africa have independently produced ecomorphs with highly similar colour patterns and (mouth) morphologies (Kocher *et al.* 1993). Here, we explore the ecological and genetic basis of one of the particular trophic structures of cichlids, which has evolved convergently in various cichlid assemblages: fleshy lips. Instead of focusing on species with hypertrophied lips between the radiations in the East African lakes, we compare the thick-lipped phenotype between a cichlid assemblage in East African (Lake Tanganyika) and in Central American (the lake Nicaragua/Managua system), where thick-lipped species have evolved in parallel (see Fig. 1).

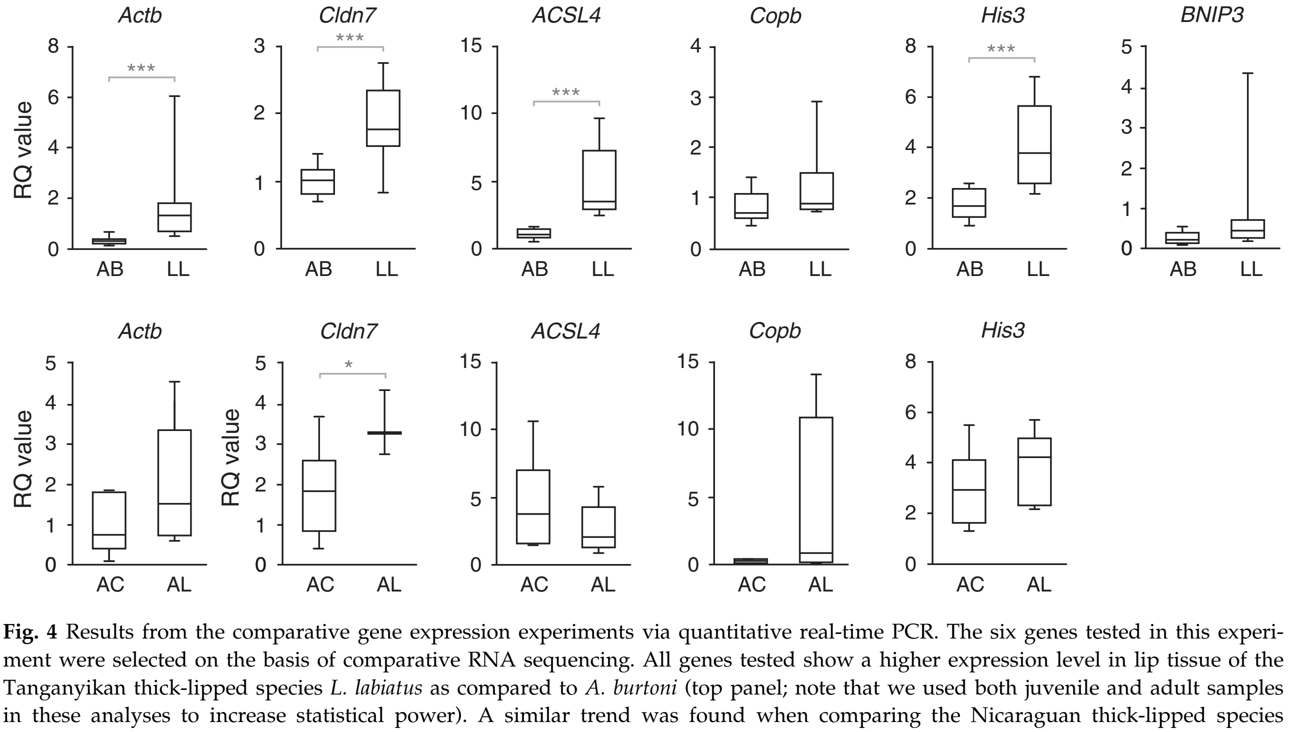


Fig. 4 Results from the comparative gene expression experiments via quantitative real-time PCR. The six genes tested in this experiment were selected on the basis of comparative RNA sequencing. All genes tested show a higher expression level in lip tissue of the Tanganyikan thick-lipped species *L. labiatus* as compared to *A. burtoni* (top panel; note that we used both juvenile and adult samples in these analyses to increase statistical power). A similar trend was found when comparing the Nicaraguan thick-lipped species *A. labiatus* to its sister species *A. citrinellus* (with the exception of *ACSL4*; lower panel). Note that *BNIP3* could not amplified in the *Amphilophus* species. *Astatotilapia burtoni* (AB); *Lobochilotes labiatus* (LL); *Amphilophus citrinellus* (AC); *Amphilophus labiatus* (AL); *$P < 0.05$; ***$P < 0.01$.

The evolution of hypertrophied lips in cichlid adaptive radiations

Our phylogenetic and demographic analyses in the Tanganyikan Tropheini and the Nicaragua Midas Cichlid species complex reveal that the thick-lipped species are nested within their respective clade. The molecular phylogeny of 14 Tropheini species (Fig. 2a) shows a footprint characteristic for adaptive radiations: a 'bottom heavy' topology with only limited phylogenetic resolution at the deeper nodes due to rapid lineage formation (Gavrilets & Vose 2005). Our new analysis thus confirms previous results based on mtDNA only (Sturmbauer *et al.* 2003) or a combination of mtDNA and AFLPs (Koblmuller *et al.* 2010). In all analyses thus far, the thick-lipped species *L. labiatus* forms an independent evolutionary lineage that branches off deep in the Tropheini. Its exact position remains unclear, though. In the AFLP phylogeny of Koblmuller *et al.* (2010), *L. labiatus* appears as sister group to all Tropheini except for the genus *Tropheus*, which is sister to all other representatives of that clade (the topology has very little support, though). In our new phylogeny and the previous mtDNA trees of Sturmbauer *et al.* (2003), *L. labiatus* shows affinities to *Simochromis* and *Petrochromis* (with moderate support). In all phylogenies, however, *L. labiatus* is nested within a clade formed by various species that feed on algae and biocover (see our character state reconstruction in Fig. 2a).

In the Midas Cichlid species complex from Central America, a phylogenetic approach is not applicable with the available molecular markers. There is simply too little genetic variation, even in the rapidly evolving mitochondrial control region, as a consequence of the young age of the assemblage (see Barluenga & Meyer 2004, 2010; Barluenga *et al.* 2006). The structures of our haplotype genealogies, which now also include the analysis of a nuclear gene (Fig. 2b), confirm this scenario. In combination with the mismatch analyses (Fig. 2c), these data suggest that *A. labiatus* underwent its main demographic expansion soon after the expansion of the sympatric *A. citrinellus* populations (see Barluenga & Meyer 2010 for a large-scale analysis of the Midas Cichlid species complex).

In both species assemblages, the evolution of the thick-lipped phenotype was associated with similar modifications of overall body shape (Fig. 3a,b). Reduced body depth, a more acute head shape and a larger mouth, along with the prominently enlarged lips, can be hypothesized to be adaptations to the species' microhabitat and trophic niche. If individuals search for food in narrow rock crevices, these modifications appear advantageous. Klingenberg *et al.* (2003) already suggested that the elongation of the head, as observed in both '*labiatus*' species, increases suction power. Other morphological differences between the two thick-lipped species, such as eye size or the length of anal fin insertion, might be either due to adaptations to the specific environments or due to phylogenetic effects. Inclusion of other thick-lipped species in future studies focusing on the ecology and morphological evolution of this trait might answer this question.

The function of hypertrophied lips in cichlids

Hypertrophied lips in cichlids have been implicated in several functions. For example, it has been suggested that fleshy lips are used to seal cracks and grooves to facilitate sucking of invertebrates (Barlow & Munsey 1976; Ribbink *et al.* 1983; Seehausen 1996; Konings 1998), that they act as bumpers to protect from mechanical shock (Greenwood 1974; Yamaoka 1997) or that they function as taste (Arnegard *et al.* 2001) or mechanoreceptors (Fryer 1959; Fryer & Iles 1972). Previous food web analyses on *L. labiatus* identified this species as mollusc eater (Nori 1997).

Our ecomorphological analysis of the thick-lipped species *L. labiatus* from Lake Tanganyika and *A. labiatus* from the large lakes in Nicaragua suggests that this phenotype is indeed associated with feeding on hard-shelled prey such as snails, mussels and crustaceans in rocky habitats (Fig. 3c). We cannot, however, conclusively answer the question whether the lips are used to seal rock crevices or whether they serve as bumpers or receptors. In the underwater observations at our field site at Lake Tanganyika, small snails were usually engulfed by *L. labiatus* via suction feeding, whereas larger snails were turned around several times before being swallowed or spit out (see Appendix S4). This would classify the lips as instrument to handle hard-shelled invertebrate food (mostly molluscs). Note, however, that our observations were made in semi-natural conditions only, in the form of concrete ponds equipped with stones from the lake and filled with lake water.

Our experimental set-up could not address the possibility that phenotypic plasticity plays a role in the formation of fleshy lips, as has previously been shown in certain foraging traits in cichlid fishes (oral jaws: Meyer 1987; pharyngeal jaws: e.g. Greenwood 1965; Huysseune 1995; Muschick *et al.* 2011). Interestingly, it has been reported that thick-lipped cichlid species lose their fleshy lips under unnatural conditions in captivity (when fed with standard food; Barlow & Munsey 1976; Barlow 1976; Loiselle 1998). So far, there is no evidence for the opposite process, the plastic development of fleshy lips due to environmental or feeding properties. In the common garden experiment of Muschick *et al.*

(2011), one group of normally lipped *A. citrinellus* individuals was fed with whole snails over a period of several months, and—although not formally assessed—no increase in lip size was apparent (compared to the other two treatment groups peeled snails and crushed snails). Another study on a snail crusher (Huysseune 1995) did not report such changes either, which seems to suggest that phenotypic plasticity in the lips, if at all present, is specific to thick-lipped species only. Future common garden and feeding experiments should thus expand on this question. Such experiments, combined with molecular analyses, should focus on the plastic component of this trait and its genomic basis.

Insights into the genetic basis of hypertrophied lips in cichlids

Our comparative gene expression assays with RNA sequencing between tissue from thick and normal lips identified a set of 141 candidate genes that might be responsible for the morphogenesis or the maintenance of fleshy lips in (East African) cichlid fish (Appendix S3). Six genes were tested further by means of quantitative real-time PCR, and these experiments largely confirm the results obtained from RNAseq (Fig. 4). While there is no obvious functional connection to fleshy lips for three of these differentially expressed genes (*ACSL4*, *His3* and *BNIP3*), the observed upregulation of the remaining three (*Actb*, *Cldn7* and *Copb*) makes sense in the light of the structure of hypertrophied lips. These three genes (together with *BNIP3*) also show a higher expression in lip tissue from adult vs. juvenile *L. labiatus* (Appendix S5).

It has previously been shown that the 'fleshy' lips of the Lake Malawi cichlid *Otopharynx pachycheilus* mainly consist of loose connective tissue covered by dermis and a layer of epithelial cells (Arnegard *et al.* 2001). Interestingly, the known functions of *Actb*, *Cldn7* and *Copb* can be directly implicated in cell and/or intercell or membrane structure. The cytoplasmic *Actb* is found in high abundance in nonmuscle cells, where it promotes cell surface and cell thickness (Schevzov *et al.* 1992), which is also consistent with its upregulation in the more massive adult compared to juvenile *L. labiatus* lips (Appendix S5). The integral membrane protein *Cldn7* (among other *claudin* gene family members) constitutes the backbone of tide junctions between epithelial cells (Tsukita *et al.* 2001). The coatomer coat proteins (such as *Copb*) are involved in protein and membrane trafficking via vesicle secreting between the endoplasmic reticulum and the Golgi apparatus, plus the intra-Golgi transport (Duden 2003). In addition, they mediate lipid homoeostasis and lipid storage for energy use and membrane assembly (Soni *et al.* 2009). *Copb* might thus be involved in cellular (membrane) development but possibly also in the formation of fat cells that compose adipose tissue, a specific subtype of connective tissue. Clearly, much more work will be necessary to unravel the development and genetic basis of hypertrophied lips in cichlids, for which we herewith established a valuable starting ground.

Our results, especially the comparison of gene expression levels between the thick-lipped species in East Africa and Central America (Fig. 4), allow us to touch on ongoing discussions related to the genetic basis of convergent morphologies (reviewed in Brakefield 2006; Arendt & Reznick 2008; Elmer & Meyer 2011). Although our qPCR results in Midas Cichlid (*Amphilophus* spp.) species must be taken with caution (efficiency was lower as a consequence of using molecular tools developed for the African species leading to a lack of statistical power), we find rather similar trends in gene expression. Our results seem to indicate that a largely overlapping set of genes was recruited to develop the hypertrophied lips in Nicaraguan and African species, which are—according to most authors—separated by ~ 100 million years of evolution. This important question about the basis of convergent phenotypes should be addressed in future studies, and thick-lipped fish species, including those outside the family Cichlidae, appear as an excellent model system.

Conclusion

Our integrative evolutionary, ecological, morphological, observational and genomic analysis of thick-lipped species in East Africa and in Nicaragua reveals stunning similarities between these convergent morphs. Both thick-lipped species appear to have evolved early in the respective clade, they seem to have adapted to the same habitat (rocks) and food source (hard-shelled prey), and their evolution was associated with comparable morphological trajectories, especially in the mouth and head region. Importantly, we also show that the expression patterns of at least some genes are similar, too. We thus provide valuable resource for future studies focusing on the development of this trait and genetic basis of convergence.

Acknowledgements

We would like to thank our helpers in the field, V. Campos, C. Heule, B. Meyer, M. Roesti; E. P. van den Berghe and G. Tembo and his crew for their logistic support in Nicaragua and Africa, respectively; the Ministerio del Ambiente y los Recursos Naturales Nicaragua (MARENA) and the Lake Tanganyika Research Unit, Department of Fisheries, Republic of Zambia, for research permits; and three anonymous referees and the Subject Editor, C. Eizaguirre, for valuable comments. This study was supported by grants from the Fundacao para Ciencia e a Tecnologia (FCT, Portugal) to M. E. S., the Swiss Academy of Sciences to A. I., the Spanish Minsiterio de Economía y Competitividad to M. B., and the European Research Council (ERC, Starting Grant 'INTERGENADAPT'), the University of Basel and the Swiss National Science Foundation (SNF, grants 3100A0_122458 and CRSII3_136293) to W. S.

References

Anders S, Huber W (2010) Differential expression analysis for sequence count data. *Genome Biology*, **11**, R106.

Arendt J, Reznick D (2008) Convergence and parallelism reconsidered: what have we learned about the genetics of adaptation? *Trends in Ecology and Evolution*, **23**, 26–32.

Arnegard ME, Snoeks J, Schaefer SA (2001) New three-spotted cichlid species with hypertrophied lips (Teleostei: Cichlidae) from the deep waters of Lake Malaŵi/Nyasa, Africa. *Copeia*, **2001**, 705–717.

Baldo L, Santos ME, Salzburger W (2011) Comparative transcriptomics of Eastern African cichlid fishes shows signs of positive selection and a large contribution of untranslated regions to genetic diversity. *Genome Biology and Evolution*, **3**, 443–455.

Barlow GW (1976) The Midas cichlid in Nicaragua. In: *Investigations of the Ichthyofauna of Nicaraguan lakes* (ed. Thorson TB), pp. 333–358. School of Life Sciences, University of Nebraska, Lincoln, Nebraska.

Barlow GW, Munsey JW (1976) The Red Devil Midas Cichlid species complex in Nicaragua. In: *Investigations of the Ichthyofauna of Nicaraguan Lakes* (ed. Thorson TB), pp. 359–370. School of Life Sciences, University of Nebraska, Lincoln, Nebraska.

Barluenga M, Meyer A (2004) The Midas cichlid species complex: incipient sympatric speciation in Nicaraguan cichlid fishes? *Molecular Ecology*, **13**, 2061–2076.

Barluenga M, Meyer A (2010) Phylogeography, colonization and population history of the Midas cichlid species complex (*Amphilophus* spp.) in the Nicaraguan crater lakes. *BMC Evolutionary Biology*, **10**, 326.

Barluenga M, Stolting KN, Salzburger W, Muschick M, Meyer A (2006) Sympatric speciation in Nicaraguan crater lake cichlid fish. *Nature*, **439**, 719–723.

Berner D, Roesti M, Hendry AP, Salzburger W (2010) Constraints on speciation suggested by comparing lake-stream stickleback divergence across two continents. *Molecular Ecology*, **19**, 4963–4978.

Blankenberg D, Von Kuster G, Coraor N (2010) Galaxy: a web-based genome analysis tool for experimentalists. *Current protocols in molecular biology/edited by Frederick M. Ausubel et al.* Chapter 19, Unit 19 10 11-21.

Brakefield PM (2006) Evo-devo and constraints on selection. *Trends in Ecology and Evolution*, **21**, 362–368.

Brichard P (1989) Cichlids and all Other Fishes of Lake Tanganyika, T.H.F. Publications, Neptune City, New Jersey.

Bruford MW, Hanotte O, Brookfield JFY, Burke T (1998) Multilocus and single-locus DNA fingerprinting. In: *Molecular Analysis of Populations* (ed. Hoelzel AR), pp. 283–336. Oxford University Press, New York.

Butler MA, Sawyer SA, Losos JB (2007) Sexual dimorphism and adaptive radiation in Anolis lizards. *Nature*, **447**, 202–205.

Diepeveen ET, Salzburger W (2011) Molecular characterization of two endothelin pathways in East African cichlid fishes. *Journal of Molecular Evolution*, **73**, 355–368.

Duden R (2003) ER-to-Golgi transport: COP I and COP II function (Review). *Molecular Membrane Biology*, **20**, 197–207.

Elmer KR, Meyer A (2011) Adaptation in the age of ecological genomics: insights from parallelism and convergence. *Trends in Ecology and Evolution*, **26**, 298–306.

Excoffier L, Laval G, Schneider S (2005) Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online*, **1**, 47–50.

Fryer G (1959) The trophic interrelationships and ecology of some littoral communities of Lake Nyasa with a special reference to the fishes, and a discussion on the evolution of a group of rock-frequenting Cichlidae. *Proceedings of the Zoological Society of London*, **132**, 153–281.

Fryer G, Iles TD (1972) The Cichlid Fishes of the Great Lakes of Africa: Their Biology and Evolution. Oliver & Boyd, Edinburgh.

Gavrilets S, Losos JB (2009) Adaptive radiation: contrasting theory with data. *Science*, **323**, 732–737.

Gavrilets S, Vose A (2005) Dynamic patterns of adaptive radiation. *Proceeding of the National Academy of Sciences U S A*, **102**, 18040–18045.

Genner MJ, Seehausen O, Lunt DH *et al.* (2007) Age of cichlids: new dates for ancient lake fish radiations. *Molecular Biology and Evolution*, **24**, 1269–1282.

Giardine B, Riemer C, Hardison RC *et al.* (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Research*, **15**, 1451–1455.

Goecks J, Nekrutenko A, Taylor J (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, **11**, R86.

de Graaf M, Dejen E, Osse JWM, Sibbing FA (2008) Adaptive radiation of Lake Tana's (Ethiopia) *Labeobarbus* species flock (Pisces, Cyprinidae). *Marine and Freshwater Research*, **59**, 391–407.

Grant PR, Grant BR (2008) How and Why Species Multiply: The Radiations of Darwin's Finches. Princeton University Press, Princeton, USA.

Greenwood PH (1965) Environmental effects on the pharyngeal mill of a cichlid fish, *Astatoreochromis alluaudi*, and their taxonomic implications. *Proceedings of the Linnean Society London*, **176**, 1–10.

Greenwood PH (1974) The cichlid fishes of Lake Victoria East Africa: the biology and evolution of a species flock. *Bulletin of the British Museum for Natural History (Zool.) Supplementary*, **6**, 1–134.

Harmon LJ, Kolbe JJ, Cheverud JM, Losos JB (2005) Convergence and the multidimensional niche. *Evolution*, **59**, 409–421.

Herder F, Schwarzer J, Pfaender J, Hadiaty RK, Schliewen UK (2006) Preliminary checklist of sailfin silversides (Teleostei: Telmatherinidae) in the Malili Lakes of Sulawesi (Indonesia), with a synopsis of systematics and threats. *Verhandlungen der Gesellschaft für Ichthyologie*, **5**, 139–163.

Huysseune A (1995) Phenotypic plasticity in the lower pharyngeal jaw dentition of *Astatoreochromis alluaudi* (Teleostei: Cichlidae). *Archives in Oral Biology*, **40**, 1005–1014.

Johnson MA, Revell LJ, Losos JB (2009) Behavioral convergence and adaptive radiation: effects of habitat use on territorial behavior in Anolis lizards. *Evolution*, **64**, 1151–1159.

Katoh K, Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics*, **9**, 286–298.

Klingenberg CP (2011) MorphoJ: an integrated software package for geometric morphometrics. *Molecular Ecology Resources*, **11**, 353–357.

Klingenberg CP, Barluenga M, Meyer A (2003) Body shape variation in cichlid fishes of the *Amphilophus citrinellus* species complex. *Biological Journal of the Linnean Society*, **80**, 397–408.

Koblmuller S, Egger B, Sturmbauer C, Sefc KM (2010) Rapid radiation, ancient incomplete lineage sorting and ancient hybridization in the endemic Lake Tanganyika cichlid tribe Tropheini. *Molecular Phylogenetics and Evolution*, **55**, 318–334.

Kocher TD (2004) Adaptive evolution and explosive speciation: the cichlid fish model. *Nature Reviews Genetics*, **5**, 288–298.

Kocher TD, Conroy JA, McKaye KR, Stauffer JR (1993) Similar morphologies of cichlid fish in lakes Tanganyika and Malawi are due to convergence. *Molecular Phylogenetics and Evolution*, **2**, 158–165.

Konings A (1998) Tanganyikan Cichlids in their Natural Habitat. Cichlid Press, El Paso.

Lang M, Miyake T, Braasch I *et al.* (2006) A BAC library of the East African haplochromine cichlid fish *Astatotilapia burtoni*. *Journal of Experimental Zoology. Part B, Molecular and Developmental Evolution*, **306B**, 35–44.

Lee WJ, Conroy J, Howell WH, Kocher TD (1995) Structure and evolution of teleost mitochondrial control regions. *Journal of Molecular Evolution*, **41**, 54–66.

Li H, Handsaker B, Wysoker A *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Loiselle PV (1998) The *Amphilophus labiatus* Species Complex. The Cichlid Room Companion. Retrieved on June 05, 2012, from: http://www.cichlidae.com/article.php?id=106.

Losos JB (2009) Lizards in an Evolutionary Tree: Ecology and Adaptive Radiation of Anoles. University of California Press, Berkeley.

Losos JB (2010) Adaptive radiation, ecological opportunity, and evolutionary determinism. American Society of Naturalists E. O. Wilson award address. *American Naturalist*, **175**, 623–639.

Losos JB (2011) Convergence, adaptation and constraint. *Evolution*, **65**, 1827–1840.

Losos JB, Ricklefs RE (2009) Adaptation and diversification on islands. *Nature*, **457**, 830–836.

Losos JB, Jackmann TR, Larson A, De Queiroz K, Rodrigues-Schettino L (1998) Contingency and determinism in replicated adaptive radiations of island lizards. *Science*, **279**, 2115–2118.

Matschiner M, Hanel R, Salzburger W (2011) On the origin and trigger of the notothenioid adaptive radiation. *PLoS One*, **6**, e18911.

Meyer A (1987) Phenotypic plasticity and heterochrony in *Cichlasoma managuense* (Pisces, Cichlidae) and their implications for speciation in cichlid fishes. *Evolution*, **41**, 1357–1369.

Meyer A, Morrissey JM, Schartl M (1994) Recurrent origin of a sexually selected trait in *Xiphophorus* fishes inferred from a molecular phylogeny. *Nature*, **368**, 539–542.

Muschick M, Barluenga M, Salzburger W, Meyer A (2011) Adaptive phenotypic plasticity in the Midas cichlid fish pharyngeal jaw and its relevance in adaptive radiation. *BMC Evolutionary Biology*, **11**, 116.

Nori M (1997) Structure of littoral fish communities. In: *Fish Communities in Lake Tanganyika* (eds Kawanabe H, Hori M, Nagoshi M), pp. 277–298. Kyoto University Press, Kyoto.

Nosil P, Crespi BJ, Sandoval DP (2002) Host-plant adaptation drives the parallel evolution of reproductive isolation. *Nature*, **417**, 440–443.

Posada D (2008) jModelTest: phylogenetic model averaging. *Molecular Biology and Evolution*, **25**, 1253–1256.

Ribbink AJ, Marsch BA, Marsch AC, Ribbink AC, Sharp BJ (1983) A preliminary survey of the cichlid fishes of rocky habitats in Lake Malawi. *South African Journal of Zoology*, **18**, 149–310.

Roesti M, Hendry AP, Salzburger W, Berner D (2012) Genome divergence during evolutionary diversification as revealed in replicate lake–stream stickleback population pairs. *Molecular Ecology*, **21**, 2852–2862.

Rohlf FJ (2006) DIG, Version 2.10.0. Department of Ecology and Evolution. State University of New York, Stony Brook, NY.

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.

Rundle HD, Nagel L, Wenrick Boughman J, Schluter D (2000) Natural selection and parallel speciation in sympatric sticklebacks. *Science*, **287**, 306–308.

Salzburger W (2009) The interaction of sexually and naturally selected traits in the adaptive radiations of cichlid fishes. *Molecular Ecology*, **18**, 169–185.

Salzburger W, Meyer A (2004) The species flocks of East African cichlid fishes: recent advances in molecular phylogenetics and population genetics. *Naturwissenschaften*, **91**, 277–290.

Salzburger W, Meyer A, Baric S, Verheyen E, Sturmbauer C (2002) Phylogeny of the Lake Tanganyika cichlid species flock and its relationship to the Central and East African haplochromine cichlid fish faunas. *Systematic Biology*, **51**, 113–135.

Salzburger W, Mack T, Verheyen E, Meyer A (2005) Out of Tanganyika: Genesis, explosive speciation, key-innovations and phylogeography of the haplochromine cichlid fishes. *BMC Evolutionary Biology*, **5**, 17.

Salzburger W, Renn SC, Steinke D, Braasch I, Hofmann HA, Meyer A (2008) Annotation of expressed sequence tags for the East African cichlid fish *Astatotilapia burtoni* and evolutionary analyses of cichlid ORFs. *BMC Genomics*, **9**, 96.

Salzburger W, Ewing GB, Von Haeseler A (2011) The performance of phylogenetic algorithms in estimating haplotype genealogies with migration. *Molecular Ecology*, **20**, 1952–1963.

Schevzov G, Lloyd C, Gunning P (1992) High level expression of transfected beta- and gamma-actin genes differentially impacts on myoblast cytoarchitecture. *The Journal of Cell Biology*, **117**, 775–785.

Schluter D (2000) The Ecology of Adaptive Radiation. Oxford University Press, New York.

Schluter D, Nagel LM (1995) Parallel speciation by natural selection. *American Naturalist*, **146**, 292–301.

Schoener TW (1970) Nonsynchronous spatial overlap of lizards in patchy habitats. *Ecology*, **51**, 408–418.

Seehausen O (1996) Lake Victoria Rock Cichlids, Verdujin Cichlids, Zevenhuizen, The Netherlands.

Seehausen O (2006) African cichlid fish: a model system in adaptive radiation research. *Proceeding of the Royal Society London B*, **273**, 1987–1998.

Sereno PC, Wilson JA, Conrad JL (2004) New dinosaurs link southern landmasses in the Mid-Cretaceous. *Proceeding of the Royal Society London B*, **271**, 1325–1330.

Sibbing FA, Nagelkerke LAJ, Stet RJM, Osse JWM (1998) Speciation of endemic Lake Tana barbs (*Cyprinidae*, Ethiopia) driven by trophic resource partitioning: a molecular and ecomorphological approach. *Aquatic Ecology*, **32**, 217–227.

Simpson GG (1953) The Major Features of Evolution, Columbia University Press, New York.

Soni KG, Mardones GA, Sougrat R, Smirnova E, Jackson CL, Bonifacino JS (2009) Coatomer-dependent protein delivery to lipid droplets. *Journal of Cell Science*, **122**, 1834–1841.

Stiassny MLJ, Meyer A (1999) Cichlids of the Rift Lakes. *Scientific American*, **280**, 64–69.

Sturmbauer C, Hainz U, Baric S, Verheyen E, Salzburger W (2003) Evolution of the tribe Tropheini from Lake Tanganyika: synchronized explosive speciation producing multiple evolutionary parallelism. *Hydrobiologia*, **500**, 51–64.

Swofford DL (2003) PAUP*—Phylogenetic Analyses Using Parsimony and Other Methods, Version 4.0, Sinauer, Sunderland, Massachusetts.

Tsukita S, Furuse M, Itoh M (2001) Multifunctional strands in tight junctions. *Nature Reviews in Molecular Cell Biology*, **2**, 285–293.

Verheyen E, Salzburger W, Snoeks J, Meyer A (2003) Origin of the superflock of cichlid fishes from Lake Victoria, East Africa. *Science*, **300**, 325–329.

Villa J (1982) Peces Nicaragüenses de Agua Dulce, Fondo de Promoción Cultural, Banco de América, Managua.

Wake DB, Wake MH, Specht CD (2011) Homoplasy: from detecting pattern to determining process and mechanism of evolution. *Science*, **331**, 1032–1035.

Wallace RK (1981) An assessment of diet-overlap indexes. *Transactions of the American Fisheries Society*, **110**, 72–76.

Yamaoka K (1997) Trophic ecomorphology of Tanganyikan cichlids. In: *Fish Communities in Lake Tanganyika* (eds Kawanabe H, Hori M, Nagoshi M), pp. 27–56. Kyoto University Press, Kyoto.

Yoder JB, Clancey E, Des Roches S *et al*. (2010) Ecological opportunity and the origin of adaptive radiations. *Journal of Evolutionary Biology*, **23**, 1581–1596.

M.C., E.T.D., M.E.S. and A.I. are PhD students in the group of W.S. M.C. is interested in parallel evolution events as natural replicates to test hypotheses about trait evolution and the different (or similar) genetic bases that underlie these phenotypes. E.T.D. is interested in the genetic basis of adaptive traits and the selective forces acting upon these genes. M.E.S. is interested in the ecological and developmental mechanisms underlying the emergence and diversification of novel adaptive traits. A.I. is interested in ecomorphological adaptations, phylogeography and taxonomy in cichlid fishes. M.M. recently finished his PhD in the group of W.S. and is now postdoctoral fellow with Patrik Nosil in Sheffied. His research is concerned with morphological and genomic evolution in adaptive radiations. N.B. is a technical assistant who is involved in several projects of the SalzburgerLab. M.B. is a group leader at the Natural History Museum in Madrid. Her research focuses on understanding incipient stages of speciation and the sequence of adaptations and specializations that organisms undergo after the colonization of new habitats. W.S. is Professor of Zoology and Evolutionary Biology at the University of Basel. The research of his team focuses on the genetic basis of adaptation, evolutionary innovation and animal diversification. The main model systems in the laboratory are threespine stickleback fish, Antarctic notothenioids and the exceptionally diverse assemblages of cichlid fishes. The laboratory's homepage at http://www.evolution.unibas.ch/salzburger/ provides further details on the group's (research) activities.

Data accessibility

Newly generated DNA sequences for phylogenetic and haplotype analyses have been deposited in GenBank under accession numbers JX402217–JX402407 (see Appendix S1 for details). Illumina reads from the RNA-seq experiments are available from the Sequence Read Archive (SRA) at NCBI under the accession number SRA052992. Data from the stomach and gut content analyses, the MorphoJ input files and the quantitative real-time PCR experiments have been deposited at Dryad (doi:10.5061/dryad.vf1ms).

Supporting information

Additional Supporting Information may be found in the online version of this article.

Appendix S1 List of specimens used in this study including sampling date and location and GenBank accession numbers.

Appendix S2 PCA of overall body shape of the African cichlid *Lobochilotes labiatus* and the Nicaraguan species *Amphilophus labiatus* and *A. citrinellus* (a) and distribution of landmarks for morphometric analyses (b).

Appendix S3 Blast2GO annotations of genes with differential expression between lip tissue from thick-lipped and normal-lipped cichlid species.

Appendix S4 Underwater video showing snail feeding in *Lobochilotes labiatus*.

Appendix S5 Results of the quantitative real-time PCR experiments comparing adult and juvenile lip tissue of the African cichlid species *Lobochilotes labiatus* and *Astatotilapia burtoni*.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Chapter 7

Discussion and future perspectives

Discussion and future perspectives

The main goal of this doctoral thesis was to broaden our understanding of the genetics and developmental basis of the emergence and diversification of egg-spots, an evolutionary novelty in East African cichlid fishes. This work has led to the discovery of many genes that play a role in the adult egg-spot, and has given some important insights into the genetics of emergence of this trait. Here, I review the main results and conclusions of the experiments, and analyze their contribution to the topic of evolution of new morphological traits.

With the development of next generation sequencing techniques, it is now relatively easy to perform expression profiling in a specific tissue in order to identify candidate genes. Chapter two describes a study in which we qualitatively described the transcriptomes of two important species for our research. This analysis revealed that the genetic distance between cichlids is fairly small (1.33%) and that there is a large contribution of untranslated regions (0.38%) to the genetic diversity. We also developed a database of genes that underwent Darwinian positive selection and that possibly underlie some of the amazing adaptations seen in cichlid fishes. The availability of these transcriptomes represents a major resource of molecular tools for studying genetics in cichlids, and revealed itself immensely useful for this thesis.

Egg-spots are a novel sexually dimorphic trait present only in haplochromine cichlid fishes [1]. One of the exciting results of chapter three is the involvement of a duplicate pair or genes, *fhl2a*/*fhl2b*, in the development of the egg-spot, and the correlation of a SINE insertion upstream of *fhl2b* with the emergence of this trait. Moreover, this SINE insertion, in conjunction with the first intron of *fhl2b*, has enhancer properties and preliminary results suggest that this region might be a male biased enhancer, driving expression in pigment cells and fin ray segments. In tetrapods, FHL2 is known to interact with the androgen receptor (AR) which is a major effector of secondary sexual differentiation by activating or inhibiting gene expression [2][3]. The interaction of androgen-receptor with tissue/cell specific transcription factors, or co-factors (eg. *fhl2*), activates further downstream targets that will in turn activate the sexual differentiation cascade in a tissue specific manner [4]. The change in the regulatory region of *fhl2b* in the ancestor of the modern haplochromines might have recruited this gene to both pigment cells and fin segments, where it activates further downstream targets together with AR. Interestingly, several AR binding sites were found in the SINE element insertion (data not shown). The observation that *fhl2b* upstream region (including the SINE element and intron 1) might be a male biased enhancer further strengthens this hypothesis of AR and FHL2B interacting to form egg-spots. AR might enhance *fhl2b* gene expression, which in turn acts together with AR to activate further downstream targets that will contribute to the pigmentation and egg-spot development in the fin. Clearly, these hypotheses have to be tested with functional assays, which unfortunately are not easily available for cichlids. Other research groups have reported progress on this topic [5], and we definitely must try and reproduce their transgenesis in our laboratory species, *A. burtoni*, in order to determine the role of *fhl2a* and *fhl2b*. Overall, our results suggest that egg-spot emergence involved co-option of a gene after a gene duplication event. This gene encodes a transcriptional co-factor that possibly affects downstream protein-protein interactions and therefore alters gene expression programs. Together, our results suggest that gene duplication and regulatory changes, that induce new downstream protein-protein interactions, play a role in the development of the novel egg-spot trait.

Chapter four is a study that aimed to increase our knowledge of possible genes underlying the egg-spot phenotype. To this end, we conducted a description of the egg-spot transcriptome by comparing fins within individuals. Many of these genes were confirmed as egg-spot genes and we generated hypotheses about their possible function using inter-species gene expression comparisons. We found evidence for co-option of patterning candidates (e.g. HoxC12a) that are correlated with the presence of the egg-spot and we should further test the functionality of these genes. An interesting outcome of this work was the realization that there are cichlid specific transcripts which seem to be correlated with the presence of egg-spots. This suggests that not only co-option of pre-existing genes/networks, but also that the recruitment of lineage specific genes is involved in the emergence of novel traits. This result joins the accumulating evidence that lineage specific genes are more important than once thought and that more emphasis should be put on their study [6, 7]. With this chapter we were only able to correlates egg-spot presence with gene expression. In the future we shall correlate the expression of these candidate genes with the development of the egg-spot in juvenile *A.burtoni* fish as well as extending our studies to other species. Multispecies comparisons of gene expression through development will show how genes interact and how their interactions shape the variability seen in this phenotype.

With chapters three and four I have established the egg-spot as a model trait for the study of novelties and generated many candidates that will be useful for future studies. This thesis focused primarily on the genes underlying the trait, and not so much in the precise developmental processes that are involved in its morphogenesis. The next step is to describe in detail the morphogenesis of the egg-spot, including a full characterization of the pigmentation cells involved, their migration movements and interaction behaviours. We also need to understand if there is a landmark on the anal fin where egg-spots first start to develop. Observations indicate that, in *A. burtoni*, the first egg-spot always appears on the fifth fin ray but that the rest of the egg-spot pattern is variable (personal observation). It is of extreme importance to follow up on this observation and determine the dynamics of development and the dynamics of concomitant gene expression. Only by comparing these dynamics we will be able to determine the role that our candidate genes play on the morphogenesis of the trait. Further comparison and integration of gene expression and developmental processes in other species with different phenotypes will shed light on the evolution of the trait and give us insight into how intra- and inter-variability is shaped.

Chapter five is a perspective on cichlids as a model system and stressed an important point: integration of data is needed in order to fully understand the evolution of traits. In order to understand novelty we need: 1) good phylogenetic framework where we can map our traits and dissect its evolutionary history; 2) a good phenotypic and developmental characterization of the trait; 3) closely related species where we can exert the comparative method in order to understand how changes in genes parallel changes in phenotypes and finally 4) we need to be able to experimentally test our findings with gene functional assays. Concerning our egg-spot project we definitely need to improve points 2 and 4. The advance in the cichlid model system was astonishing in the last four years, especially considering genome and transgenesis availability, these advances will allow cichlids and egg-spots to be established as one of the premier vertebrate models to study the emergence of evolutionary novelties. In chapter six we studied parallel evolution in thick-lipped cichlids, demonstrating the variety of questions that cichlids can be used to address, especially considering the genetic resources now available.

This thesis has contributed substantially to the understanding of the egg-spots emergence and diversification, and can be used as a basis for a range of exciting studies into the origin and diversification of novel traits, as well as in cichlid diversification in general. We have an extensive list of candidate genes and insights into one of the possible mechanisms that contributed to its origin – re-deployment of a gene via mobile element insertion in a *cis*- regulatory region.

References

1. Salzburger W, Mack T, Verheyen E, Meyer A: **Out of Tanganyika: genesis, explosive speciation, key-innovations and phylogeography of the haplochromine cichlid fishes.** *BMC Evolutionary Biology* 2005, **5**:17.

2. Müller JM, Isele U, Metzger E, Rempel A, Moser M, Pscherer A, Breyer T, Holubarsch C, Buettner R, Schüle R: **FHL2, a novel tissue-specific coactivator of the androgen receptor.** *The EMBO Journal* 2000, **19**:359-369.

3. Borg B: **Androgens in teleost fishes**. *Comparative Biochemistry and Physiology* 1994, **109C**:219-245.

4. Williams TM, Carroll SB: **Genetic and molecular insights into the development and evolution of sexual dimorphism.** *Nature Reviews Genetics* 2009, **10**:797-804.

5. Fujimura K, Kocher TD: **Tol2-mediated transgenesis in tilapia (Oreochromis niloticus).** *Aquaculture* 2011, **319**:342-346.

6. Nowick K, Stubbs L: **Lineage-specific transcription factors and the evolution of gene regulatory networks.** *Briefings in Functional Genomics* 2010, **9**:65-78.

7. Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TCG: **More than just orphans: are taxonomically-restricted genes important in evolution?** *Trends in Genetics* 2009, **25**:404-13.

Acknowledgments

First, I would like to thank Prof. Walter Salzburger for giving me the opportunity to perform my PhD studies in his group and for the guidance and support throughout these years.

I acknowledge Fundação para a Ciência e Tecnologia for providing funding support to this project and my fellowship.

I would also like to thank Dr. Ingo Braasch for support, discussions and excellent collaboration, without which the functional assays from this PhD thesis would not be possible.

I am immensely grateful to Nicolas Boileau for both his professional support and friendship, it has been invaluable throughout.

I must express my special gratitude to Hugo Gante and Matthew Hall for insightful discussions and support during the more difficult times of my PhD. They were a great source of motivation.

I would like to thank Dr. Élio Sucena for the inspiration that finally convinced me to pursue a career in Evolutionary Biology, for the help whilst writing my PhD grant proposal and for current manuscript commentaries.

I owe thanks to all members of the Salzburger lab, both past and present. Too many to name in detail here. In particular I would like to thank my very special office mates, in particular Anya, Britta and Adrian for their unconditional support and fun times throughout.

A special thanks to Brigitte Aeschbach and Astrid Boehne for being incredibly helpful in the lab.

My masters student Romina provided a year of work for this project, along with an amazing friendship.

Julien Veziliér must be thanked for showing me the research paper that led me do my PhD studies in the Salzburger lab.

Fabio Cortesi, Moritz Muschick and David Duneau have provided essential friendship and PhD companionship throughout.

To my flatmates and dear friends, Frauke Münzel, Alexandre Barras and Michael Wagner who have provided me with an essential escape and made my time as a PhD student in Basel very special indeed.

Finally, I owe a very special thanks to Ian Warren, who provided invaluable help, support and kept me balanced.

M. Emilia Santos

Born on 14/11/84 in Machico, Portugal

155 Rue Cuvier, 69006 Lyon, France

T +33 7 81 72 11 15
emilia.p.santos@gmail.com

Research Experience and Education

Current: Postdoctoral Research Associate, Genomics of Development and Evolution research group, Ecole Normale Supérieure de Lyon, France (PI: Dr. Abderrahman Khila)

2013: Postdoctoral Research Associate, Social Evolution and Social Behaviour research group, University of Bristol, UK (6months) (PI: Dr. Seirian Sumner)

2012-2013: Postdoctoral Research Associate, Salzburger lab, University of Basel, Switzerland (PI: Walter Salzburger)

2008-2012: PhD degree in Evolutionary Biology, University of Basel, supervised by Prof. Walter Salzburger; final mark: summa cum laude.

2006-2008: Master degree in "Evolutionary biology and development"; Faculdade de Ciências da Universidade de Lisboa; final mark: 18 (0-20).

2002–2006: Degree (Licenciatura) in Biology; Faculdade de Ciências da Universidade de Lisboa; final mark 16 (0-20).

Publications

Santos ME, Braasch I, Boileau N, Meyer B, Boehne A, Affolter M, Salzburger W (2014) The evolution of cichlid egg-spots are linked with a *cis*-regulatory change. *Nature Communications*, doi:10.1038/ncomms6149

Brawand D, Wagner CE, Li YI, Malinsky M, Keller I, Fan S, Simakov O, Ng AY, Lim ZH, Bezault E, Turner-Maier J, Johnson J, Alcazar R, Noh HJ, Russell P, Aken B, Alföldi J, Amemiya C, Azzouzi N, Baroiller JF, Barloy-Hubler F, Berlin A, Bloomquist R, Carleton KL, Conte MA, D'Cotta H, Eshel O, Gaffney L, Galibert F, Gante HF, Gnerre S, Greuter L, Guyon R, Haddad NS, Haerty W, Harris RM, Hofmann HA, Hourlier T, Hulata G, Jaffe DB, Lara M, Lee AP, MacCallum I, Mwaiko S, Nikaido M, Nishihara H, Ozouf-Costaz C, Penman DJ, Przybylski D, Rakotomanga M, Renn SCP, Ribeiro FJ, Ron M, Salzburger W, Sanchez-Pulido L, **Santos ME**, Searle S, Sharpe T, Swofford R, Tan FJ, Williams L, Young S, Yin S, Okada N, Kocher TD, Miska EA, Lander ES, Venkatesh B, Fernald RD, Meyer A, Ponting CP, Streelman JT, Lindblad-Toh K, Seehausen O & Di Palma F (2014) The genomic substract for adaptive radiation in African cichlid fish. *Nature*, doi:10.1038/nature13726

Colombo M, Diepeveen ET, Muschick M, **Santos ME,** Indermaur A, Boileau N, Barluenga M, Salzburger W (2013) The ecological and genetic basis of convergent thick-lipped phenotypes in cichlid fishes. *Molecular Ecology*, 22:670-684.

Santos ME, Salzburger W (2012) How cichlids diversify. *Science*, 338:619-621.

Baldo L, **Santos ME**, Salzburger W (2011) Comparative transcriptomics of Eastern African cichlid fishes shows signs of positive selection and a large contribution of untranslated regions to genetic diversity. *Genome Biology and Evolution*, 3:443-55.

Santos ME, Athanasiadis A, Leitão AB, Dupasquier L, Sucena E (2011) Alternative splicing and gene duplication in the evolution of the FoxP gene sub-family. *Molecular Biology and Evolution*, 28(1):237-47.

Lopes PC, Sucena E, **Santos ME**, Magalhães S (2008) Rapid experimental evolution of pesticide resistance in C. elegans entails no costs and affects the mating system. *PLoS ONE*, 3(11): e3741.

Further publications

Santos ME, Baldo L, Gu X, Salzburger W (in preparation) Transcriptomics of a novel and variable pigment trait in cichlid fishes - identification of candidate genes for egg-spot development.

Grants

2013: Marie-Curie Intra-European fellowship for career development (Initiative FP7-PEOPLE-2013-IEF) for 24 months (194'046.60 EUR)

2013: SNSF (Swiss National Science Foundation) Early Postdoc Mobility fellowship for 18months (64,000.00 CHF)

2012: Society for Molecular Biology and Evolution, Travel Award (700 USD)

2009: University of Basel, “Evolutionary Biology in Guarda” workshop (280 CHF)

2008: FCT (Fundação para a Ciência e Tecnologia , Portuguese Science Foundation) PhD fellowship (SFRH/BD/43421/2008) for 4 years (130'000.00 EUR)

2006/2007: IEFP (Instituto de Emprego e Formação Profissional), Fellowship for a 9 month internship for the project “Alternative splicing and gene duplication in the evolution in the Foxp gene sub-family” at IGC (Instituto Gulbenkian the Ciência)

Scientific training course attendance

2011: Workshop on Analysis of Differential Gene Expression, Swiss Institute of Bioinformatics, Lausanne, Switzerland

2011: Workshop on Comparative Genomics, Ceski Krumlov, Czech Republic

2009: Evolutionary Biology in Guarda 2010 – Workshop from the University of Basel and ETH Zurich – Guarda, Switzerland

2008: Real time PCR training – Course from Applied Biosystems – Advancing Quality Science, November, Basel Switzerland.

2007: Molecular Evolution, Phylogenetics and Adaptation. Course from the Gulbenkian Training Programme in Bioinformatics, Oeiras, Portugal.

2006: Evolution course integrated in the PhD Program of the Instituto Gulbenkian de Ciência, October, Oeiras, Portugal.

2006: Genetics course integrated in the PhD Program of the Instituto Gulbenkian de Ciência, October, Oeiras, Portugal.

2003: PADI Diving course – Open water diver, Funchal, Portugal.

Presentations at scientific conferences

2014: Trancriptome profiling of a key morphological innovation: the propelling fan of the water walking bug *Rhagovelia obesa"* Euro Evo Devo (EED) 2014, Vienna, Austria

2014: "Trancriptome profiling of a key morphological innovation: the propelling fan of the water walking bug *Rhagovelia obesa"* Society for Molecular Biology and Evolution (SMBE) 2014, San Juan, Puerto Rico

2013: "The molecular basis of a novel pigment trait in cichlid fishes" 14th meeting of European Society for Evolutionary Biology (ESEB), Lisbon, Portugal

2013: "The molecular basis of a novel pigment trait in cichlid fishes" Biology 13, University of Basel, Switzerland

2012: "The genetic basis of an evolutionary novelty in cichlid fishes." VIII Encontro Nacional de Biologia Evolutiva (ENBE) 2012, Instituto Gulbenkian de Ciência, Oeiras, Portugal.

2012: "The genetic basis of an evolutionary novelty in cichlid fishes." Cichlid Science 2012, University of Leuven, Belgium.

2012: "Spot the difference! The genetic basis of an evolutionary novelty in cichlid fishes." Society for Molecular Biology and Evolution (SMBE) 2012, Dublin

2011: "Genetic basis of parallel evolution in egg-spots of cichlid fishes" 13th meeting of European Society for Evolutionary Biology (ESEB), Tübingen, Germany

2011: "Genetic basis of parallel evolution in egg-spots of cichlid fishes" 17th meeting of European Meeting of PhD Students in Evolutionary Biology (EMPSEB) Seia, Portugal

2011: "Genetic basis of parallel evolution in egg-spots of cichlid fishes" Evolution 2011, Oklahoma, United States of America

2010: "Genetic basis of parallel evolution in egg-spots of cichlid fishes." 16th EMPSEB, Wierzba, Poland.

2010: "Genetic basis of a colour trait in cichlid fishes." Cichlid Science 2010, University of Basel, Switzerland.

Invited Research Seminars

2013: "The molecular basis of a novel pigment trait in cichlid fishes." Research seminar, IGFL, France.

2010: "Alternative splicing and gene duplication in the evolution of the FoxP gene subfamily." Research seminar, University of Basel, Switzerland.

2010: "Genetic basis of a colour trait in cichlid fishes." University of Sussex, United Kingdom.

Teaching experience

2009-2010: Organization of the interaction seminar for the Zoological Institute and Masters program at the University of Basel.

2009: Teaching (tutoring) Introduction to biology (Fall course). Level: 1st year Bachelor, University of Basel. Time: 16h

2009: Supervisor of scientific projects during a 1-month course focused on undergraduate research skills (Spring course). Level: 3rd year Bachelor, University of Basel. Time: 60h.

Training and student supervision

2010-2011: Romina Celozzi "Genetic basis of egg-spot development in *Ophthalmotilapia ventralis*", Walter

Salzburger group, University of Basel, Switzerland.

2010-2011: Vasco Campos “The role of copy number variation (CNV) in Tanganyikan cichlids”, Walter Salzburger group, University of Basel, Switzerland.

2008-2010: Corina Heule “Development and morphology of egg-spots in cichlid fishes” partial supervision, Walter Salzburger group, University of Basel, Switzerland.

2008: Vincent Martin “The role of copy number variation (CNV) in Tanganyikan cichlids”, Walter Salzburger group, University of Basel, Switzerland.

Awards

Prize for second best talk at Biology 13, Basel, Switzerland

Prize for third best talk at EMPSEB 17, Seia, Portugal

Society membership

European Society for the Evolutionary Biology; Society for the Study of Evolution; Portuguese Society for Evolutionary Biology

Foreign Languages

Portuguese: Native speaker

English: Spoken: Very good; Reading: Very good; Written: Very good

French: Spoken: Good; Reading: Good; Written: Good

German: Spoken: Basics, Reading: Basics; Written; Basics

Miscellaneous

Volunteer at the Zoologic Garden of Lisbon between 2002 and 2004.

Volunteer at the Lisbon Oceanarium between 2003 and 2005.

Working for “Fun Science” (Children Science teaching) from 2004 to 2006.

Two seasons of field work with Tanganyikan cichlids in Mpulungu, Zambia (2010 and 2011).