DOCTOR OF PHILOSOPHY

Investigating mechanisms of genome expansion in Corydoradinae catfish

Marburger, Sarah

Award date:
2015

Awarding institution:
Bangor University

PRIFYSGOL
BANGOR
UNIVERSITY

NERC
SCIENCE OF THE ENVIRONMENT

Investigating Mechanisms of Genome Expansion in Corydoradinae catfish

A thesis submitted to Bangor University for the degree of Doctor of Philosophy

By Sarah Marburger, B.Sc., MRes

September 2015

Molecular Ecology and Fisheries Genetics Laboratory

School of Biological Sciences

Environment Centre Wales

Bangor University

Bangor, Gwynedd, LL57 2UW

Declaration and Consent

Details of the Work

I hereby agree to deposit the following item in the digital repository maintained by Bangor University and/or in any other repository authorized for use by Bangor University.

Author Name: ……..

Title: ……..

Supervisor/Department: ……..

Funding body (if any): ……..

Qualification/Degree obtained: ……..

This item is a product of my own research endeavours and is covered by the agreement below in which the item is referred to as "the Work". It is identical in content to that deposited in the Library, subject to point 4 below.

Non-exclusive Rights

Rights granted to the digital repository through this agreement are entirely non-exclusive. I am free to publish the Work in its present version or future versions elsewhere.

I agree that Bangor University may electronically store, copy or translate the Work to any approved medium or format for the purpose of future preservation and accessibility. Bangor University is not under any obligation to reproduce or display the Work in the same formats or resolutions in which it was originally deposited.

Bangor University Digital Repository

I understand that work deposited in the digital repository will be accessible to a wide variety of people and institutions, including automated agents and search engines via the World Wide Web.

I understand that once the Work is deposited, the item and its metadata may be incorporated into public access catalogues or services, national databases of electronic theses and dissertations such as the British Library's EThOS or any service provided by the National Library of Wales.

I understand that the Work may be made available via the National Library of Wales Online Electronic Theses Service under the declared terms and conditions of use (http://www.llgc.org.uk/index.php?id=4676). I agree that as part of this service the National Library of Wales may electronically store, copy or convert the Work to any approved medium or format for the purpose of future preservation and accessibility. The National Library of Wales is not under any obligation to reproduce or display the Work in the same formats or resolutions in which it was originally deposited.

Statement 1:

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree unless as agreed by the University for approved dual awards.

Signed …….. (candidate)

Date ……..

Statement 2:

This thesis is the result of my own investigations, except where otherwise stated. Where correction services have been used, the extent and nature of the correction is clearly marked in a footnote(s).

All other sources are acknowledged by footnotes and/or a bibliography.

Signed …….. (candidate)

Date ……..

Statement 3:

I hereby give consent for my thesis, if accepted, to be available for photocopying, for inter-library loan and for electronic storage (subject to any constraints as defined in statement 4), and for the title and summary to be made available to outside organisations.

Signed …….. (candidate)

Date ……..

Statement 4:

Choose **one** of the following options

a)	I agree to deposit an electronic copy of my thesis (the Work) in the Bangor University (BU) Institutional Digital Repository, the British Library ETHOS system, and/or in any other repository authorized for use by Bangor University and where necessary have gained the required permissions for the use of third party material.	
b)	I agree to deposit an electronic copy of my thesis (the Work) in the Bangor University (BU) Institutional Digital Repository, the British Library ETHOS system, and/or in any other repository authorized for use by Bangor University when the approved **bar on access** has been lifted.	
c)	I agree to submit my thesis (the Work) electronically via Bangor University's e-submission system, however I **opt-out** of the electronic deposit to the Bangor University (BU) Institutional Digital Repository, the British Library ETHOS system, and/or in any other repository authorized for use by Bangor University, due to lack of permissions for use of third party material.	

Options B should only be used if a bar on access has been approved by the University.

In addition to the above I also agree to the following:

1. That I am the author or have the authority of the author(s) to make this agreement and do hereby give Bangor University the right to make available the Work in the way described above.

2. That the electronic copy of the Work deposited in the digital repository and covered by this agreement, is identical in content to the paper copy of the Work deposited in the Bangor University Library, subject to point 4 below.

3. That I have exercised reasonable care to ensure that the Work is original and, to the best of my knowledge, does not breach any laws – including those relating to defamation, libel and copyright.

4. That I have, in instances where the intellectual property of other authors or copyright holders is included in the Work, and where appropriate, gained explicit permission for the inclusion of that material in the Work, and in the electronic form of the Work as accessed through the open access digital repository, *or* that I have identified and removed that material for which adequate and appropriate permission has not been obtained and which will be inaccessible via the digital repository.

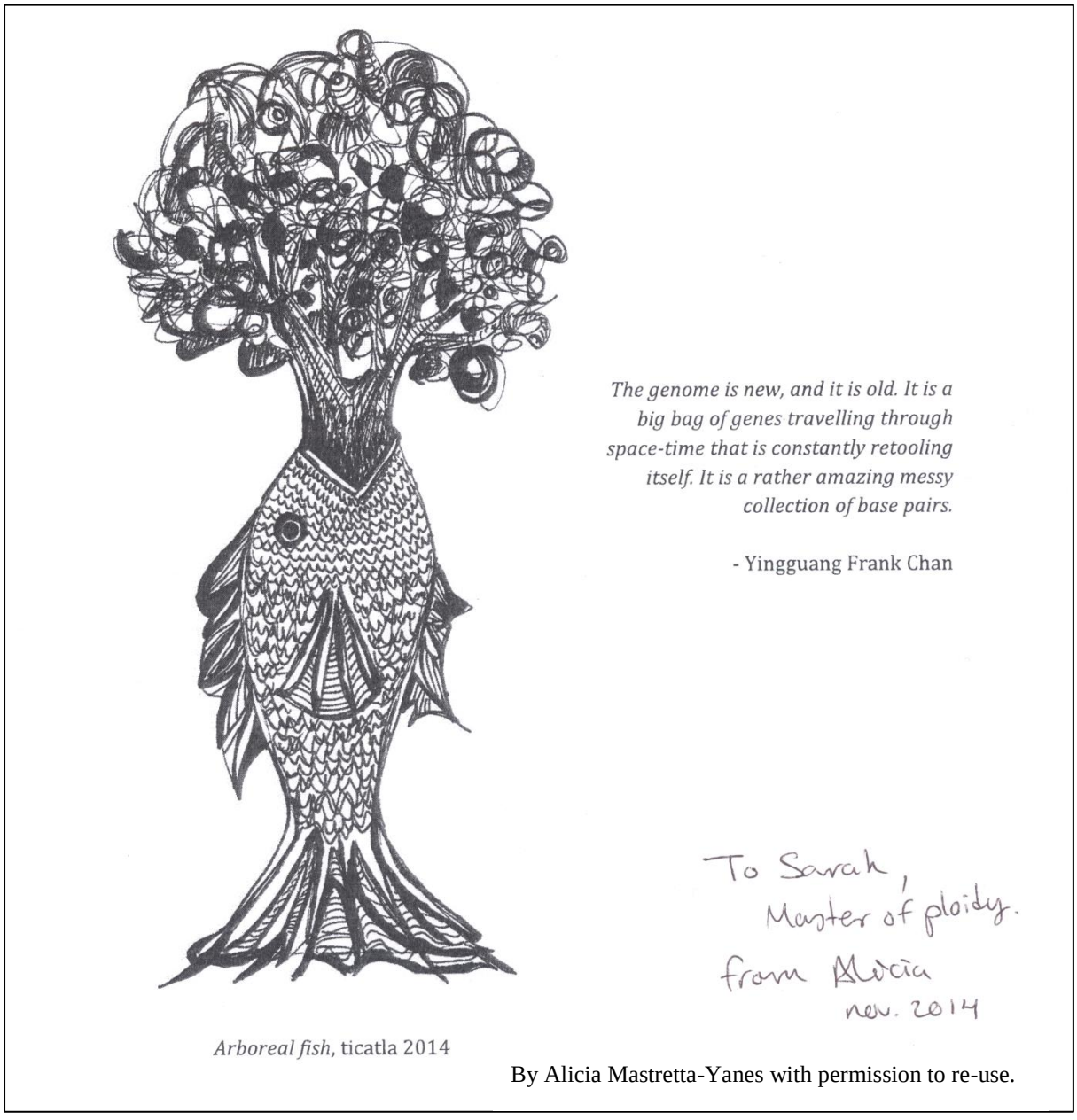
5. That Bangor University does not hold any obligation to take legal action on behalf of the Depositor, or other rights holders, in the event of a breach of intellectual property rights, or any other right, in the material deposited.

6. That I will indemnify and keep indemnified Bangor University and the National Library of Wales from and against any loss, liability, claim or damage, including without limitation any related legal fees and court costs (on a full indemnity bases), related to any breach by myself of any term of this agreement.

Signature: …… Date : ……..

Dedication

To my wonderful family, Ursula, Gerd and Marc André Marburger



By Alicia Mastretta-Yanes with permission to re-use.

Abstract

The Corydoradinae catfish are a diverse sub-family of neo-tropical catfishes (order Siluriformes) with more than 170 species described to date. One of the most compelling features of this sub-family is the enormous amount of variation in genome size. With species containing between 0.5 pg and 4.8pg of DNA, variation is comparable to that found across the Teleostei as a whole. Previous phylogenetic analysis identified nine distinct lineages within the Corydoradinae, with more basal lineages possessing smaller genomes and with largest genome sizes found in the most derived lineages. To date, nothing is known about the mechanism that drove this genome expansion in the Corydoradinae, though Whole Genome Duplication (WGD) events have been suggested. Here, the incidence of WGD events has been investigated using a *Hox* gene and a restriction site associated DNA (RAD) sequencing data set. Both data sets identified a major duplication event at the base of the group, with additional duplication events occurring across the family. These duplication events were shown to have led to relaxed purifying selection and increased functional divergence of *HoxA13a* copies in the Corydoradinae compared with teleosts that have not undergone additional rounds of WGD. The RAD data set confirmed significant genome-wide shifts in duplicate, multi-haplotype regions across the Corydoradinae, and indicates that several species from lineages 6-9 are functionally polyploid, whereas species that underwent earlier WGDs have largely diploidized and are likely paleopolyploids. An increase in paralogous genes was noted, with Gene Ontology suggesting that gene retention in the Corydoradinae mirrors previously described retention in *Tetraodon* following the fish-specific genome duplication in the Teleostei. Intriguingly, the RAD data also identified a significant expansion of Transposable Elements (TEs), driven by a DNA TE superfamily (*Tc1-Mariner*). This expansion significantly contributed to the genome size variation, though to a lesser degree than the WGD events identified within this thesis.

Acknowledgements

These past four years have been an absolutely incredible journey and there are so many people that contributed to this work in many different ways. Most of all, I owe my fantastic supervisors, who have gone above and beyond (particularly these last few weeks!), an enormous amount of gratitude. Martin must be one the most patient and kind PhD supervisors out there. Not only has he been a continuous source of ideas and patient advise (even when faced with the most naïve of questions), but also a gentle guide along what has been a very steep learning curve! I am very grateful also for him giving me the option to join him in his new lab at UEA. Si and Gary have always remained highly involved with this project, and continuously offered advice. Their help and input, particularly in the last few weeks, has been invaluable and I cannot thank them enough for proof-reading chapter after chapter while on travel or leave and while faced with many other urgent deadlines or work that needed their attention. Dr John Taggart at Sterling University has taught me most of what I know about RAD sequencing, and provided incredible support in the lab, with materials and with ideas and discussions. Without his help, this PhD would not be what it is now.

Many other people at Bangor deserve a mention, particularly Wendy Grail, who has helped with work while in Bangor and from who I have learned many tricks in the lab! Hazel and Iliana, you two are the best thesis-companions ever! Thanks for staying in touch. And thanks Hazel for helping out with printing and forms that need to be handed in in Bangor and for keeping me calm! Good luck to both of you for your future adventures, keep going! Also on the moral-support front, a shout out to the Geek-Squad (Anna, Lorena, Sonia, Robert, Dan, Tom et al) for many wine-fuelled evenings of entertainment in and around Bangor, miss you guys!

The last two years of this PhD project were spent in Norwich at UEA, and again I was lucky and met an amazing bunch of awesome people that have provided excellent moral support no matter the circumstance. A very, very special and heart-felt thank you goes to Ents, Ellen, Kat and Mr Collins, who have been spectacular friends and supported me through some bad patches (as well as through many awesome ones!!). Dave & Laura….will you still play board games with me now that I am finally free again?? Hehe thanks for all the cake and fun, here is to much more! Will, you crazy man, keep going. Thanks for all the moral support and fourth year comfort hugs.

Thanks is also due to my wonderful family (Mama, Papa, Marc und Opa) who have supported/put up with my crazy British university adventure with much love, patience and support and without whom I would never be where I am now. Hab euch lieb, ihr seid die Besten! Then there is my wonderful boyfriend Dan, who has been the most incredible support system throughout the last few years (and while finishing his very own PhD-thesis!). I cannot believe that we have both come to the end of our journey now. Thanks also to his lovely parents, Jenny and Kevin, who have been providing me with a 'British home' and much love and support over these last two and a bit years. It meant more than you can imagine.

Finally, I would like to thank NERC for funding this PhD project (grant number NE/J500203/1).

Table of Contents

List of Figures

List of Tables

Chapter 1 -Mechanisms of Genome Expansion in Corydoradinae Catfish – An Introduction

1.1 Mechanisms of Genome Size Variation

Ever since the mid-20th century when DNA (and not protein) was found to be the 'hereditary material' of which genes are formed, scientists have been fascinated with enormous variation in DNA content across the tree of life (Gregory 2005a). Initially, it was assumed that organisms with more DNA must as a consequence also possess more genes, and that more complex species therefore contain more DNA. However, it became clear very quickly that there was no correlation between the amount of DNA (genome size) and organismal complexity. This became known as the C-value paradox (C-value being a measure of haploid DNA content) (e.g. Mirsky & Ris 1951; Thomas 1971, reviewed in Gregory 2005). Today we know that the majority of eukaryotic genomes consist of mostly non-coding DNA, which explains why the amount of DNA does not correspond to the number of genes. As such, the C-value paradox is solved, but is replaced with what Ryan Gregory coined the 'C-value enigma', which highlights that while the C-value paradox itself may be answered, there is still plenty we do not understand about the mechanisms underpinning genome size variation and the subsequent phenotypic and evolutionary implications (Gregory 2001; Gregory 2005a; Gregory 2005b).

Genome size varies 200,000 fold across eukaryotes, and we still do not fully understand the mechanisms underpinning this extreme variation (Gregory 2001). Mechanisms that can lead to increases in genome size include increases in the size of introns (intergenic regions), chromosome-level duplications, tandem duplications as well as Transposable Element expansions. Whole Genome Duplication (WGD) events, also known as polyploidization, can lead to an instantaneous increase in DNA content. This is technically speaking, however, not an increase in genome size per se, but rather an increase in the number of genomes present within an organism.

Outstanding questions include whether there are patterns in the variation of genome size across taxa? What are the contributions of different mechanisms to this variation within the genome and what are the functions or triggers? What impact do these mechanisms have on an organisms phenotype, morphology and adaptive evolution? This thesis will be focusing on two mechanisms of genome size variation, namely WGD and transposable elements.

1.1.1 Polyploidy as an evolutionary mechanism

Modes of Whole Genome Duplication

Polyploidization is a well-established phenomenon that can be both ancient, occurring early in the evolutionary history of lineages, as well as an on-going process (Otto and Whitton, 2000). Polyploidy can be split into two broad categories. Firstly, polyploidization can be the product of hybridization between two distinct species involving genome merger and genome doubling, which is termed allopolyploidy. In strict allopolyploids, chromosome sets of the parental species are usually sufficiently diverged for bivalent formation during meiosis. Secondly, polyploidy may occur through doubling of the genome of a single species through unreduced gametes during meiosis, or alternatively through somatic doubling. This is known as autopolyploidy. As the homeologs are highly similar, autopolyploids usually form multivalents during meiosis, often resulting in polysomic inheritance (Stebbins, 1971, Parisod et al. 2010, Ramsey & Schemske, 2002, Otto 2007). In the case of allopolyploidy, the parental progenitors of newly formed allopolyploids may not always have sufficiently diverged for complete bivalent formation, with some chromosome pairs forming bivalents (homologs), and some forming multivalent homeologs (i.e. paralogous chromosome pairs resulting from duplication). This is referred to as segmental allopolyploidy, and it is generally well accepted that there is a continuum from doubling of identical genomes (autopolyploidy) to the doubling of highly differentiated genomes (strict allopolyploidy) (e.g. Parisod et al. 2010, Otto 2007).

Polyploidy is particularly common in plants. Estimates of the frequency differ among studies, but roughly 42% of ferns, 32% of monocots and 18% of dicots appear to have undergone a polyploidization event at some stage in their evolutionary history (Otto & Whitton 2000). In angiosperms, polyploidy appears to be a ubiquitous phenomenon and the

question is not if WGD events occurred, but how often (Soltis et al. 2009). WGD events have also occurred in fungi (Dujon et al. 2004; Ma et al. 2009; Albertin & Marullo 2012), and are particularly well studied in yeast (Morel et al. 2015; Marcet-Houben & Gabaldón 2015). While traditionally viewed as rarer in animals, it is now accepted that multiple rounds of WGD occurred at the base of all vertebrates (Dehal & Boore 2005; Putnam et al. 2008; Cañestro & Albalat 2012), and that polyploidy is a common phenomenon in amphibian and fish lineages (Mable 2004; Mable et al. 2011).

Consequences of Polyploidy

Indeed, polyploidization may be one of the most dramatic mutations known to occur within eukaryotes and may lead to genome instability, large scale genomic rearrangements and deletions, changes in gene regulation and TE activation, particularly in allopolyploids (Otto, 2007).

Though intuitively, polyploidization should be an evolutionary dead-end (e.g. inefficiency of selection when advantageous genes are masked by multiple copies, increased frequency of deleterious alleles, decreased fertility and survival through problems during meiosis) the prevalence of polyploid lineages and their evolutionary success suggests that polyploidy may have some advantages (Otto, 2007). For a new polyploid species to become established, it first needs to persist long enough for selection to be able to act upon it. While deleterious allele frequency will increase/accumulate over the long term in polyploids, deleterious effects will initially be masked, conferring new polyploids with a potential advantage. Extensive genome restructuring as a result of genome duplication (in both allo- and autopolyploids) could lead to an increase in genetic variability, with additional (redundant) copies of genes potentially exempt from functional constraints and thus a free canvas for selection to act upon. If a polyploid lineage becomes established, its long term success is very much dependent on its ability to adapt. Here again increased genetic variability may be of an advantage (Otto, 2007). Furthermore, polyploid species are less prone to inbreeding than their diploid ancestors as they will produce fewer homozygous offspring (Ronfort, 1999).

Allopolyploids have been considered more ‘successful’ polyploids as they may potentially benefit from ‘hybrid vigour’ (where hybrids have a higher fitness than either of their parental species) through increased genetic variability (Otto, 2007). Studies focusing on

flowering plants show that different modes polyploid origin can have a profound effect on gene expression: Allopolyploids often exhibit patterns of gene loss, modification of methylation patterns and nonreciprocal chromosomal exchange with effects being larger than with autopolyploids (Doyle et al., 2008). For instance, in a study that hybridized two species of *Arabidopsis*, *A. arenosa* with *A. thaliana*, 94% of the genes of *A. thaliana* were up-regulated, and the genes of *A. arenosa* correspondingly down-regulated in the hybrid (Wang *et al.*, 2006). A similar study comparing two synthetic allopolyploids in cotton revealed significantly higher expression of the maternal phenotype of *G. arboreum.* Furthermore, a study on 1400 duplicated gene pairs in cotton revealed that these biased expression ratios arose immediately as a consequence of the genomic merger. These biases can also be observed in 1-2 million year old allopolyploid cotton, making them a remarkably stable evolutionary phenomenon (Doyle et al., 2008) In addition, dosage imbalance (uneven-numbered ploidy level**)** seems to be a major factor leading to expression of novel phenotypes: E.g. in maize, triploid individuals showed a radically different expression profile in comparison to diploids. As triploids usually appear to be a necessary bridge in the formation of allopolyploids, this may also play a part in the creation of novel phenotypes (Doyle et al., 2008).

Autopolyploids have been considered both less frequent and less successful than allopolyploids, as multivalent formation during meiosis can be error prone, leading to chromosome dissegregation and thus reduced fertility (Stebbins 1971; Otto 2007; Parisod et al. 2010). However, autopolyploidy appears to be far more common and prevalent than initially thought, though it has received far less attention than allopolyploidy (Soltis & Soltis 2000; Parisod et al. 2010). Autopolyploids are characterized by genomic redundancy and polysomic inheritance, which leads to an increase in their effective population size and counteracts inbreeding depression. Genomic changes after the initial WGD event are far less dramatic than in allopolyploids with only few changes in gene expression and genomic structure, indicating that the large genomic restructuring in allopolyploids is likely the result of hybridization, more than the result of the polyploidization (Parisod et al. 2009; Parisod et al. 2010)**.** Many autopolyploids may remain undetected, as most autopolyploids show very little phenotypic divergence from their diploid progenitors. While much work is needed to address the factors may lead to successful establishment of autopolyploid lineages, evidence is accumulating that autopolyploidy coincides with periods of environmental change and may thus serve as a quick means of increasing genetic variability and escaping narrow ecological

niches (Parisod et al., 2010). In line with these predictions, a recent study shows that autopolyploid *Arabidopsis* has a higher salt tolerance compared to diploids (Chao et al. 2013).

Even though both auto- and allopolyploidy are a well-documented phenomenon across plant taxa, the evolutionary consequences of polyploidy for plant diversification are still hotly debated. For instance, it is unclear whether the frequency of polyploidy is the result of higher diversification rates of polyploid lineages, or whether it the result of frequent polyploidy formation (Mayrose et al. 2011). Polyploids appear more frequent in higher latitudes as well as in more temperate environments, which could either mean that polyploid lineages are more tolerant of stressful environments, or that these environments have led to a higher incidence of polyploidy, through unreduced gamete production for example (Moghe & Shiu 2014). Mayrose et al. (2011) compared diversification rates of polyploid angiosperms with their diploid congeners and found that polyploidy actually decreases diversification rates and maintains that polyploidy is in most cases likely to be an evolutionary dead-end. Their approach and conclusions, however, have been criticized as not statistically robust and for failing to include many prominent polyploid taxa (Soltis et al. 2014). Indeed, polyploidization may not necessarily increase diversification rates, but could lead to speciation directly through reproductive isolation from diploid parental clades (Vamosi & Dickinson 2006; Wood et al. 2009).

From Polyploid to Paleopolyploid – The process of diploidization

Before the advent of genomics, it was not appreciated just how frequently ancient polyploidization events have occurred in plant and vertebrate lineages. Following polyploidization events, genomes typically undergo extensive 'pruning' and return to an almost diploid-like state, with only traces of the ancestral duplication event remaining in the genome. This process is called diploidization, and the exact molecular mechanisms underpinning these events remain poorly understood (Wolfe 2001; Le Comber et al. 2010). Organisms that have undergone a polyploidization event in their evolutionary history but have returned to a functionally diploid state are called paleopolyploids. While new sequencing technologies allow scientists to identify more and more paleopolyploids, the process of diploidization is by no means a recent discovery. Ohno (1970) described examples of diploidization in his famous work 'Evolution by Gene Duplication', focusing in particular on Salmonidae and their gradual return from autotetraploid to diploid: in many salmonid species,

some loci show a tetrasomic inheritance pattern, whereas most loci now follow a disomic pattern. The diploidization process in Salmonidae has been considered near-complete (Ohno et al. 1970; Allendorf et al. 2015).

One of the key steps in diploidization is the return from multivalent formation to bivalent formation of chromosomes during meiosis, though the mechanisms by which this is achieved are poorly understood (Wolfe 2001). As previously mentioned, one of the biggest challenges facing newly formed polyploids is the problematic segregation of chromosomes during meiosis. This frequently leads to unbalanced gametes/aneuploidy and thus severe reduction in fertility (Comai 2005; Otto 2007; Yant et al. 2013). A reduction in chiasmata appears to be one of the mechanisms that aids stabilization of meiosis (Le Comber et al. 2010). Pairing of homologs during meiosis is under strict genetic control, and studies have demonstrated that adaptation to WGD and diploidization indeed involve selection on these genes (Cifuentes et al. 2010; Yant et al. 2013).

Diploidization is accompanied by large scale chromosomal re-arrangements and deletions leading to drastic genome downsizing (Leitch & Bennett 2004), a process which is driven by intrachromosomal recombination (Garsmeur et al. 2014) and which likely contributes to the restoration of bivalent formation (Eilam et al. 2010). The loss of duplicate regions and genes does not appear to be random. After WGD events, dosage-sensitive genes for instance are preferentially retained, a phenomenon referred to as Gene Dosage Balance Hypothesis (GDBH) (Freeling 2009). Moreover, specific groups of genes are often preferentially retained in duplicate (e.g. genes involved in DNA metabolism, or transcription factors), though these groups do not appear to be retained universally but vary with higher taxonomic differentiation (Barker et al. 2008). In allopolyploids, deletion of genetic material is often highly biased towards one sub-genome (known as biased fractionation) (Garsmeur et al. 2014). This appears to be linked to transcriptional dominance, i.e. allopolyploid species such as maize and cotton often show preferential gene expression of one dominant subgenome over the other, potentially leading to preferential loss of the non-expressed copies in the other genome (Flagel & Wendel 2010; Woodhouse et al. 2010). Currently, it is not known what drives genome dominance, though it has been proposed that differences in gene expression between sub-genomes are in part driven by differences in methylation near promoter regions (Freeling et al. 2012). This could also explain why there does not appear to be any sub-genome bias in strict autotetraploids, as sub-genomes are identical. Furthermore, random pairing of homeologs as opposed to preferential pairing may prevent the preferential deletion

of duplicate genes in a sub-genome, as random pairing could otherwise lead to individuals lacking an entire gene complement (Garsmeur et al. 2014).

1.1.2 The role of Whole Genome in early vertebrate evolution

It is widely accepted that two rounds of whole genome duplication (WGD) occurred during the early diversification of vertebrates (termed 1R and 2R duplications) (Ohno et al. 1970; Furlong & Holland 2002; Braasch et al. 2009; Dehal & Boore 2005). There is also additional strong evidence for a third fish specific genome duplication (FSGD) in ray-finned fishes that occurred before the radiation of the teleosts with subsequent lineage specific WGDs also occurring in several teleost lineages (Amores et al. 1998; J S Taylor et al. 2001; Taylor et al. 2003; Meyer & Van de Peer 2005; Kasahara et al. 2007). Fish are the most species rich group of vertebrates with approximately ~32,000 described species and possibly ~ 64,000 species in total (including undescribed species), they make up half of all vertebrate species (Nelson 1994; Glasauer & Neuhauss 2014; Froese & Pauly 2015). The impressive radiation of teleosts and the impressive diversity in regards to ecology, morphology, life history, behaviour and physiology has been associated with evolutionary implications of the FSGD (Comber and Smith, 2004; Santini *et al.*, 2009). Polyploidy appears to be more common in more basal than more derived teleosts and despite its frequent occurrence and prevalence, its role in fish evolution is still poorly understood (Leggatt and Iwama, 2004).

In vertebrates, *Hox* genes have long served as a model gene family for reconstructing the evolutionary history of lineages and were provided important evidence for the existence of early vertebrate WGD events (Amores, 1998; Chiu et al., 2002; Chiu et al., 2004; Crow et al., 2006; Yuan et al., 2010). *Hox* genes are a well-studied family of transcription factors, which play a crucial role in the early development of the anterior-posterior axis of bilateral embryos (e.g. Keynes and Krumlauf, 1994). *Hox* genes are highly structured and form clusters: Mammals possess four *Hox* clusters named A, B, C and D (Garcia-Fernàndez and Holland, 1994), while teleost fishes possess at least seven *Hox* clusters (Amores et al., 1998). *Amphioxus*, the closest relative of vertebrates, possesses only one *Hox* cluster which contains 14 genes (although this may not be the ancestral condition) (Garcia-Fernàndez and Holland, 1994). The duplication of *Hox* clusters in vertebrates and fish has been thought to be one of the major factors that lead to the increase in complexity and diversity after several rounds of

WGD in chordate evolution. Wagner and Lynch (2010) argue that *Hox* clusters in vertebrates are structurally much more constrained than invertebrate *Hox* clusters. This constraint may be relaxed directly after genome duplication, which would open a window of evolvability. However, there is no evidence for a direct correlation between *Hox* cluster duplication and evolution of complexity. For instance, invertebrates have only one *Hox* cluster while possessing by far largest variety of body plans. Both Sarcopterygians and cartilaginous fishes possess the same number of *Hox* clusters, yet, sarcopterygians are far more complex than cartilaginous fishes (Crow and Wagner, 2005). Donoghue and Purnell (2005) point out that only living taxa are considered in vertebrate evolution. When one however, also considered extinct taxa in the fossil record, the apparent causal relationship between genome duplication and evolution of complexity and diversity in vertebrates vanishes. Furthermore, a comparative study examining the diversity of ray-finned fishes in more detail found that the species richness of the teleosts is associated with two major groups, the Ostariophysi (ca 6508 species) and the Perciformes (ca. 9 293 species), both of which appeared 150-250 MY after the estimated time of the FSGD (Santini et al., 2009).

However, just because there is no evidence for a tight correlation between increased diversification rate and genome duplication does not mean the FSGD did not contribute to diversification in the long run; genome duplication is an event while diversification is a long-term process (Wagner and Lynch, 2010). Adaptation requires ecological necessity, that is, while genome duplication my contribute to opportunities for genome evolution, it does not initiate it (Crow and Wagner, 2005; Van de Peer et al., 2009). Contributing factors could include divergent resolution and/or reduced extinction risk. Divergent resolution (the loss of different copies of a duplicated gene in allopatric populations) after genome duplication may lead to an increase in speciation through reproductive isolation (Taylor et al., 2001). It has also been suggested that WGD –rather than immediately causing diversification and evolution of complexity- can decrease extinction risk, contributing to evolutionary success of these lineages in the long run. For instance genome duplication in plants appears to be clustered in time and coincide with the cretaceous tertiary boundary. It is possible that polyploidization of plants has increased their adaptability, giving them an advantage in drastically changed environments and preventing them from becoming extinct (Fawcett et al., 2009). Recently, Zhan et al (2014) analysed diversification rates of polyploid vs diploid relatives in four fish groups, the Acipenseridae (sturgeons), the Botiidae (loaches), the Cyprininae and the Salmoniformes/Esociformes. While diversification rates generally appeared higher in polyploid species, this was only significant for the Cyprininae (Zhan et al. 2014), interestingly

the group containing potentially the largest number of known polyploids in fish (Tsigenopoulos et al. 2002; Gregory & Mable 2005). Zhan et al. (2014) highlight that their analysis may suffer from sparse taxonomic sampling as well as the fact that many polyploids may simply not yet have been identified, emphasizing the need for further study.

Generally, one of the most obvious arguments as to how WGD or polyploidization could lead to diversification or adaptation in the long term is the increase in genetic material: gene and genome duplication provide more genetic material for evolution to act upon, for instance through mutation, drift and selection (Crow and Wagner, 2005). The most common fate of duplicated genes is mutational inactivation and loss (Lynch & Conery 2000). However, it appears that specific groups of genes in vertebrates are preferentially retained after having been duplicated (Doyle et al. 2008; Freeling 2009). Early genomic sequences of model fish species such as zebrafish (*Danio rerio*) and medaka (*Oryzias latipes*) indicated that gene families in fish contain more members than they do in mammals (Wittbrodt et al. 1998), speculated to be the result of an additional FSGD. We now know that roughly 20% of genes have been retained in duplicate (Braasch & Postlethwait 2012; Garcia de la Serrana et al. 2014). One group of genes that appears enriched in teleost fishes as the result of the FSGD is the pigmentary system. Braasch et al. (2009) showed that teleosts have 30% more pigmentation genes than tetrapods as a result of the FSGD which could explain the particular phenotypic teleost colour pattern and pigmentation pattern diversity.

1.1.3. How to identify Polyploids and Paleopolyploids

Traditionally, polyploids have been identified by observing duplication of chromosomal complements in closely related populations or species. A recently suggested likelihood-modelling approach may aid robust identification of ploidy events if detailed chromosome information of closely related species is available (Mayrose et al. 2010). Another certain give away that a species is polyploid is the observation of multivalents during meiosis, or indirectly the observation of for instance tetrasomic inheritance patterns for particular genes. However, recent allopolyploids may display disomic inheritance (Ohno et al. 1970; Stebbins 1971).

Paleopolyploids by their very nature are much more difficult to identify, as they are functionally diploid. The increase in gene number for instance has initially been interpreted as

an indicator of an ancient WGD event in vertebrates and teleosts, though an increase in genes itself can also be achieved through individual tandem or chromosomal duplications and thus, the existence of WGD events in vertebrates and particularly in teleosts was hotly debated until recently. More convincing evidence for a WGD event as opposed to large scale gene duplications in teleosts stems from phylogenetic and comparative mapping approaches: duplicated genes appear to have originated at the same time, show conserved synteny (i..e. gene order) and co-exist in duplicated blocks on chromosomes (Taylor et al. 2001; Vandepoele et al. 2004; Taylor et al. 2003).

A similar approach based on mapping and gene synteny led to the identification of ancient WGD events in *Arabidopsis thaliana* (Blanc et al. 2000; The Arabidopsis Genome Initiative 2000). In addition, age distributions of duplicate genes have been used extensively in the identification of paleopolyploids, that is, if a WGD event as opposed to individual gene duplication events occurred, all gene duplicates should roughly be of the same age. This technique has successfully been used to demonstrate paleopolyploidy for species such as soybean (*Glycine max*), rice (*Oryza sativa*) and maize (*Zea mays*) (Blanc & Wolfe 2004; Shoemaker et al. 2006).

The availability of genome sequences and sequencing technology has clearly greatly aided the discovery of paleopolyploid events in many species, while it is very likely that many paleopolyploids have yet to be identified, particularly in animals.

1.1.4 Transposable Elements

Transposable Elements (TEs) are loosely defined as elements that can be re-inserted into different regions within a genome. They were first described in maize (*Z. mays*) by Barbara McClintock, who discovered that the change between a stable and unstable recessive allele controlling kernel pigmentation was caused by the insertion of a TE-element (McClintock 1951). TE classification is mainly based on transposition mode. Class I elements, also known as retrotransposons, transpose via a RNA intermediate using a ‘copy-and-paste’ mechanism, whereas Class II elements, or DNA transposons, use a DNA intermediate (or in some cases no intermediate) via a ‘cut-and-paste’ mechanism (Finnegan 1989; Wicker et al. 2007). Class I elements are transcribed into RNA, which is then reverse transcribed into DNA prior to being re-integrated to a new location within the genome; the

original copy remains intact. Class II elements encode a transposase enzyme that excises the TE-element, which is then re-inserted into a target site. If this process occurs in a newly replicated DNA-strand and the new insertion site has not yet been replicated, the DNA copy will effectively contain both the original and the re-inserted TE-element (Kidwell 2005). While transposition processes and preferential insertion sites are not yet well understood for many species and TE-families, a general pattern is emerging. One type preferentially inserts distantly from genes, in AT rich regions with minimal recombination activity, potentially thus aiding escape from inactivation. Other TE-families insert within introns or near genes which tend to be GC rich (Kidwell & Lisch 1997; Kidwell 2005).

TE-elements can lead to increases in genome size

TEs are a powerful contributor to genome size and genome evolution in eukaryotes (Kidwell & Lisch 2000; Kidwell 2002; Kazazian 2004). In plants, TEs make up between 3% and 85% of genome size, with this variation mainly driven by LTR (Long Terminal Repeat) retrotransposons (Lee & Kim 2014). The overall abundance of transposable elements is less well documented outside of the plant kingdom, though TEs are known to make up a significant proportion of the genome in salamanders (Sun et al. 2012), humans (Consortium 2001) and mice (Biémont & Vieira 2006). Examples of almost all TE families have been identified in fish, making fish genomes more diverse in terms of TEs than mammals even though TEs appear far less abundant (Aparicio et al. 2002; Volff 2005).

Consequences of TE-transposition within the genome

TE activity can have tremendous impacts on genome function, and have mainly been considered highly deleterious due to their ability to insert and interrupt gene activity or gene regulation (e.g. see Kidwell & Lisch 2000; Nekrutenko & Li 2001; Kidwell 2005). Early work has largely focused on their deleterious ability, and referred to TEs as “purely selfish” or “junk”, spreading through the genome like a “cancer” (Orgel & Crick 1980; Doolittle & Sapienza 1980). Indeed, the debate on TE spread and function within genomes is still contested: the initial spread of TE elements within an organism may be underlying selfish selection, though their spread and presence within an organisms means TEs will be affected by selection at the organismal and species level (Brunet & Doolittle 2015).

In addition to interruption of gene function, a continuous increase in genome size can also be maladaptive. Increased genome size can for instance lead to increased cell sizes, changes in metabolism, and increased developmental time spans (Gregory 2005a). It is thus not surprising that TE activity is tightly controlled within the host-genome. Many TEs are suppressed through cytosine methylation, which prevents transcription. In addition, already silenced TEs are transcribed and broken into small interfering RNA (siRNA), which then act as a homology sensor for similar, potentially active TEs, and mark these for silencing (Dernburg & Karpen 2002; Aravin et al. 2007; Kim & Zilberman 2014; Kidwell 2005). It is not yet understood how similar TEs have to be in order to be successfully targeted by siRNA mechanisms. It is clear that the molecular pathways responsible for silencing new TEs within the genome differ to those from pathways that maintain silencing mechanisms. While several models have been proposed and described, further research is needed to test how these pathways work together (Fultz et al. 2015) .

Despite potential deleterious consequences and tight regulation through the host genome, TEs are not exclusively maladaptive. Kidwell and Lisch (2000) argue that there is a continuum from a parasitic to mutualistic relationship between TEs and the host genome. While TEs may be highly selfish when first invading and proliferating within a genome, TEs can be domesticated in the course of perhaps only a few million years (Nekrutenko & Li 2001; Kidwell & Lisch 2000; Volff 2006; Kidwell 2005). A study focusing on protein-coding genes in humans revealed that roughly 4% of genes contain domesticated TEs; out of these almost 90% likely first inserted into introns and were later recruited as new exons. This may be facilitated through splice sites contained within TEs, thus aiding recruitment into exonic regions (Nekrutenko & Li 2001). Particularly well known examples of domestication of TEs include genes involved in vertebrate immunity (Zhou et al. 2004; Chénais et al. 2012) and telomere and telomerase related functions in *Drosophila melanogaster* (Pardue & DeBaryshe 2011) and many more subject to discovery (Volff 2006). In addition to functional domestication in the long term, TEs are a remarkable source of genetic variability within the genome and can cause chromosomal rearrangements, alterations in gene expression as well as leading to exon shuffling (Kidwell & Lisch 2000; Feschotte & Pritham 2007; Fontdevila 2011; Kidwell 2005; Kidwell & Lisch 1997; Chénais et al. 2012). This potential of TEs to create genetic variation almost instantly has been implicated in many cases of adaptive evolution, such as adaptation to novel environments, stressors or environmental changes (Fontdevila 2011; Kidwell 2005; Chénais et al. 2012).

Genome Shock and the Re-activation of Transposable Elements

McClintock originally proposed that TE activity could be a response to stressors or what she referred to as 'genomic shocks' such as virus infections, hybridization or poison (McClintock 1984; Fontdevila 2011). Many recent studies demonstrate that indeed TEs can be re-activated and escape genome suppression pathways under certain conditions. Generally, two mechanisms of TE response have been described, one leading to the direct activation of a suppressed element, and the other indirectly leading to the activation through genome wide changes in methylation and RNAi-pathways (Fontdevila 2011). Direct activation of TEs in response to both biotic and abiotic stressors was found in cases where TE-promoter regions appear highly similar to genes implicated in the stress response. It is not clear whether genes involved in stress response have acquired their similarity through domestication of ancient TEs, or whether TEs have acquired these from the host genome (Fontdevila 2011). Examples include activation of *Tnt* retroelements in tobacco in response to acid treatment/wounding or fungal/viral/bacterial attack (Grandbastien 1998; Grandbastien et al. 2005), as well as for retroelements in *Drosophila* in response to heat and UV stressors (McDonald et al. 1997; Jardim et al. 2015) . Hybridization, and in particular hybridization in conjunction with polyploidy (i.e. allopolyploidy) are a well-documented cause of TE-activation, with examples of TE-proliferation in tobacco, wheat and rye (Fontdevila 2011).

1.2 Corydoradinae as a study system

The catfish order Siluriformes is an incredibly diverse order that is representative of the largest fraction of neo-tropical freshwater fish (Albert et al. 2011). At present, ca 3,600 species of catfish have been described (Froese & Pauly 2015), and it is estimated that an additional 35% still await discovery (Ota et al. 2015). The Callichthyidae (also known as armoured catfish) are the third largest family of catfish and the largest family of neotropical catfish (Ferraris 2007). One of the main characteristics of this family are the two large lateral bony plates that cover the entire body. The family is further subdivided into two subfamilies, namely the Corydoradinae (which contain 90% of species) and the Callichthyinae. The Corydoradinae are an extremely diverse group, with currently more than 170 species described and likely more awaiting discovery (Eschmeyer 2013). The subfamily contains four separate genera, namely *Corydoras*, *Aspidoras, Scleromystax* and *Brochis*. The genus *Corydoras* is by far the largest genus and by itself probably contains roughly 150 species (Eschmeyer 2013).

Corydoradinae are distributed throughout South America (with the exception of Chile) from the Chubut basin in Argentina in the south up until Trinidad in the north. Up until recently, the taxonomy of the Corydoradinae has been based almost entirely on colour patterns and morphology, with little genetic information (Alexandrou & Taylor 2011). Recent molecular analysis of 425 taxa using mitochondrial and nuclear markers established a comprehensive molecular phylogeny with strong support for nine distinct lineages (Alexandrou et al., 2011).

In addition to bony plates that cover their bodies, this extremely diverse group is well protected through sharp lockable spines and toxins secreted by axillary glands located underneath these. Many Corydoradinae co-occur in sympatry and share identical colour patterns, and thus have been identified as Müllerian mimics. 92% of described mimicry rings consist of distantly related species that do not compete for resources, thus indicating that Müllerian Mimicry is not powerful enough to overcome competition (Alexandrou et al. 2011). An example of mimicry rings as well as their phylogenetic composition is displayed in figure 1.

a

b

Figure 1. Example of two Müllerian Mimicry rings of Corydoradinae catfish. Mimicry rings usually consist of two to three distantly related species belonging to distinct lineages. a) From left to right: *Corydoras araguaiaensis*, *Corydoras maculifer* and *Corydoras C- sp*. b) From left to right: *Corydoras imitator* and *Corydoras C-121*. Images by Martin Taylor, reproduced with permission.

Karyotype and Genome Size Variation within the Corydoradinae

One of the most intriguing features of the Corydoradinae is their extreme variability in genome size. Early studies on karyotypes of Corydoradinae catfish hinted at extreme variability, identifying chromosome counts between 44 and 134 (Scheel et al. 1972). Further studies sampled additional species, with karyotype data now available for roughly thirty Corydoradinae. In these studies, it became apparent that chromosome numbers can differ sharply between populations thought to belong to the same species (Oliveira et al. 1992; Oliveira et al. 1993).

More recently, genome size (as haploid C-values) was measured for 206 species across the subfamily using Feulgen Image Densitometry (FID) by Alexandrou (2011) for

representatives of all 9 lineages. Obtained genome sizes ranged from 0.51 pg in lineage 1 to 4.8 pg in lineage 9. Lineages 1 to 3 do not appear to differ greatly in genome size, with a doubling in genome size apparent between lineages 4 to 8. The largest genome sizes are found in lineage 9. This apparent stepwise increase in genome-size is displayed in figure 2. To date, the exact mechanisms behind this enormous variation in genome size are not known, though it has been speculated that whole genome duplication events could have played a role in this subfamily (Turner et al. 1992; Oliveira et al. 1992; Oliveira et al. 1993; Gregory & Mable 2005). *Corydoras metae*, a lineage 9 species, has previously been named as having the largest teleost genome size so far recorded with 4.4pg, and the genome size of *C. semiaquilus* is with 0.51 pg only marginally bigger than that of the smallest recorded teleost genome of the pufferfishes with roughly 0.4pg of DNA (Gregory 2005a).Thus, the genome size variation within the subfamily Corydoradinae is near equal to the variation found across the teleosts in general.

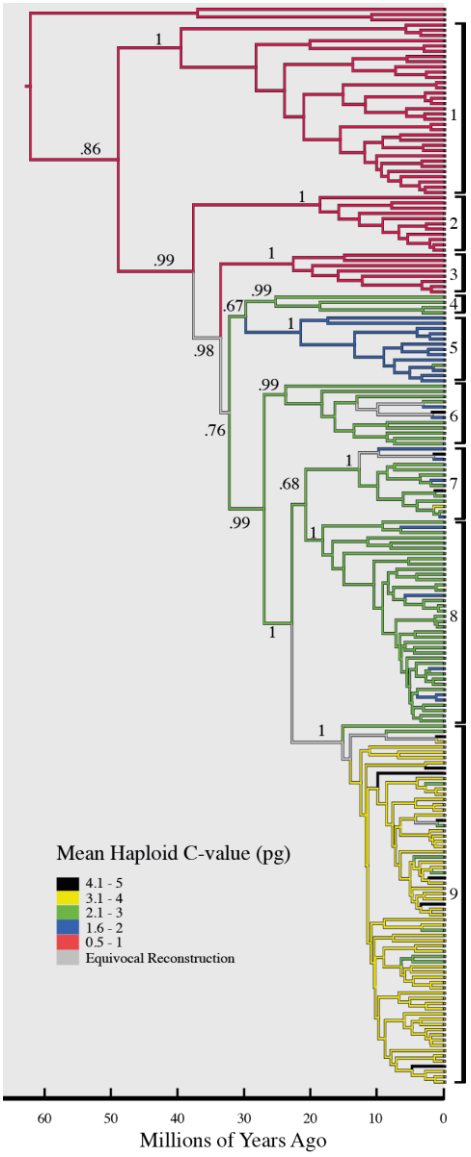


Figure 2. Phylogeny of Corydoradinae catfish (species names not shown) displaying mean C-values for each species. Modified from Alexandrou (2011). Permission to use granted by authors Markos Alexandrou and Martin Taylor.

1.3 Aims of this Thesis

The focus of this thesis is genome size expansion in the Corydoradinae catfish. Specifically, we aim to identify the underlying mechanism of the stepwise increase in DNA content across lineages. The more derived lineages in the Corydoradinae contain a higher diversity of colour patterns, more species, as well as more mimics (Alexandrou 2011). Early studies (unpublished data) also indicated that more derived lineages may potentially be more resistant to parasites. It is not clear to what extent more DNA could contribute to Corydoradinae evolution, nor what genomic consequences and re-arrangements this expansion may have caused. To fully understand whether and how the incredible genome size variation in the Corydoradinae affected their adaptive evolution, it is crucial to first identify the underlying mechanisms that drove this change.

Specifically, our objectives were to identify whether WGD events have occurred within the lineage, and if so, whether any extant species should be considered functionally polyploid. Furthermore, we wanted to quantify whether other factors could have resulted in an increase in genome size, with a particular focus on the role of TEs.

Aims of individual Chapters:

- In Chapter 1, the aim was to provide a general overview of the main mechanisms of genome size variation studied in this thesis and furthermore to discuss evolutionary consequences of these mechanisms. This chapter also introduced the model system and highlights the unanswered questions this thesis is aiming to address.
- In Chapter 2, the aim was to identify potential WGD events using *Hox* genes as a marker, as was done for teleost species in previous studies. The objectives were to identify the copy number of the *HoxA13a* gene (with the expectation that species with a higher genome size contain more alleles for this gene if WGD events have occurred), and to identify whether duplication by means of WGD events could have altered selection pressure and lead to divergent and functional evolution.
- In Chapter 3, the aim was to identify WGD across the phylogeny using a Next Generation Sequencing approach. In contrast to Chapter 2, this allowed us to identify genome wide patterns, in addition to identifying ploidy levels. By using a reduced representation approach, our objectives were to quantify haplotype number for

assembled contigs from across the genome, and to use bi-allelic SNP read ratios to identify ploidy levels.

- Chapter 4 aimed to identify genes and paralogs across the Corydoradinae using our reduced sequencing RAD data set. The objectives were to identify an increase in gene or paralog number that would be consistent with WGD events, as well as to identify differences in Gene Ontology between species, which could provide information on potential preferential gene retention.
- The aim of Chapter 5 was to identify Transposable Elements in the RAD-data set. Specifically, the objectives of this chapter were to measure the relative contribution of TEs to genome size variation in Corydoradinae in relation to WGD events identified in previous data chapters.
- Chapter 6 aims to combine and discuss findings for all data chapters, and to provide a broader context of the overall contribution and future possibilities of this research.

Chapter 2 - *HoxA13a* duplications in the Corydoradinae – evidence for multiple rounds of Whole Genome Duplication

2.1 Introduction

Originally thought of only minor importance, Ohno (1970) first postulated that Whole Genome Duplication (WGD) may have played a major role in early vertebrate evolution. While perceived as controversial at first, largely due to a lack of sufficient sequence data and rigorous methodology (e.g. Hughes et al. 2001; Martin 2001, Durand, 2003), it is now well accepted that two rounds of genome duplication (termed 1R and 2R) preceded the radiation of vertebrates between 500-800 mya (Putnam et al. 2008; Cañestro & Albalat 2012; Dehal & Boore 2005) . A further fish-specific WGD event, known as 3R, Fish-Specific Genome Duplication (FSGD) or Teleost-Specific (TS-WGD), occurred within ray-finned fishes (Actinopterygii) prior to the radiation of the teleosts, some 320-350 million years ago (Amores et al. 1998; Christoffels et al. 2004; Postlethwait et al. 2004; Hoegg et al. 2004; Vandepoele et al. 2004; Meyer & Van de Peer 2005; Crow et al. 2006). Further lineage-specific duplications in the Teleostei have been recorded for instance in the Salmonidae as well as in the Cyprinidae (Gregory & Mable 2005).

Prior to availability of genomic data, early evidence for ancient duplications was obtained from *Hox* genes, and these have since been used as a marker for whole genome duplications in vertebrates (Crow et al. 2009; Crow et al. 2012; Glasauer & Neuhauss 2014). *Hox* genes (or homeobox genes) are a set of highly conserved transcription factors that play a fundamental role in early body plan development along the anterior-posterior axis. They are linearly expressed and organized within a cluster. To our knowledge, no tandem duplications of *Hox* genes have been observed in vertebrates. Invertebrates possess only one *Hox* cluster, whereas lobe-finned fishes, amphibians, reptiles and mammals (Sarcopterygii) appear to contain four clusters termed A, B, C and D. Such events are proposed to be the product of two rounds of ancient whole genome duplications. In teleosts, seven or eight clusters are present, which are termed Aa, Ab, Ba, Bb and so forth (Amores et al. 1998; J S Taylor et al. 2001; Volff 2005; Guo et al. 2010; Henkel et al. 2012).

Ohno (1970) speculated that gene and genome duplications provide the raw material for natural selection, and saw it as a creative force of evolution that could accelerate diversification and the development of new character traits. Specifically, he proposed that after a duplication event, the duplicate gene copy would be freed from selection pressures and could potentially evolve new functions. Many workers have since speculated that the morphological diversification in vertebrates could have been facilitated by the 1R and 2R WGD events (e.g. Durand 2003; Cañestro & Albalat 2012). Similarly, the TS-WGD event has also been linked to the impressive radiation of the Teleostei (Van de Peer et al. 2009; Glasauer & Neuhauss 2014). Support comes from studies that indeed place the timing of the TS-WGD prior to the radiation of teleosts (Christoffels et al. 2004; Vandepoele et al. 2004). Teleosts are the most diverse group of vertebrates, with currently ~32,000 species described, and estimates of total species ranging up to 64,000 (Glasauer & Neuhauss 2014). However, the duplication/divergence hypothesis has been criticized, largely because of the significant time delay between the TS-WGD event and the teleost diversification (Santini et al. 2009). Other lineages that have undergone a WGD event, such as the Salmonidae, have not undergone subsequent radiation and with 228 species are a comparatively small family (Froese & Pauly 2015). It should be noted, however, that a recent study identified an increased speciation rate and morphological diversification in the Salmonidae compared to other groups (Rabosky et al. 2013). The TS-WGD most likely provided the genomic diversity that drove the complexity and diversification of the Teleostei, even though it may not have been its direct driver (Glasauer & Neuhauss 2014).

The Corydoradinae are the most diverse family of catfishes (Order: Siluriformes), with more than 170 species described to date, a minimum of 150 within the genus *Corydoras* (Eschmeyer 2013). This extremely diverse group is native to neotropical America and well protected through sharp, lockable spines and toxins secreted by axillary glands. In addition to their phenotypic diversity, Corydoradinae also exhibit marked variation in genome size, with C-values ranging from 0.51pg to 4.8pg of DNA (Alexandrou 2011). More derived species generally have higher genome sizes, with the smallest genomes occurring in the oldest lineages. It is not known what drove the genome size expansion within this family, though several rounds of WGD could explain the difference in genome size in different lineages of the family. Thus, the Corydoradinae potentially provide an excellent group to study the effects of WGD events on the evolution and diversification of vertebrates.

The aim here was to investigate whether an increase in genome size in the Corydoradinae may derive from one or several rounds of Whole Genome Duplication events, using *Hox* gene copy-number as a marker for duplication events. Primers for *HoxA13a* in the Corydoradinae were developed, and used as a marker to investigate WGD, as in Zebrafish (*Danio rerio*) (Crow et al. 2009) and American Paddlefish (Crow et al. 2012). The objectives were

1) to identify copy number of the *HoxA13a* gene in Corydoradinae species

2) to test whether duplication events could have led to changes in selection pressure on *HoxA13a* copies, which could have led to functional divergence.

2.2 Methods

Taxon Sampling and DNA extraction

Samples were wild caught between 2005 and 2013 across South America by Martin Taylor and Claudio Oliveira and were stored in absolute Ethanol. DNA was extracted from fin tissue using a salt extraction protocol after Sunnucks & Hales (1996) and Aljanabi & Martinez (1997). Problematic samples that failed to amplify were re-extracted using the Qiagen Blood and Tissue kit.

In order to detect potential WGD events across the phylogeny, species representative of all lineages of the Corydoradinae were sampled (lineages are displayed in figure 2, chapter 1). For each lineage, several species were tested and the samples that amplified most reliably were selected (with a minimum of one species per lineage). Particular emphasis was placed lineage 9, which contains the species with the largest genome sizes.

PCR Amplification and Cloning

The *HoxA13a* was initially characterised in the Corydoradinae using a degenerate primer set developed by Yuan et al. (2010) (table 1), with an expected insert size of approximately 1,198bp based on blunt snout bream (*Megalobrama amblycephala*). A single species was amplified from mtDNA lineage 1 (*C. semiaquilus)* and lineage 9 (*C. metae).* For PCR amplification, we added 0.2mM forward and reverse primers, 1.25µg/µl of bovine serum albumin (BSA, Biolabs), 0.25mM per dNTP, 1.5mM $MgCl_2$ (Promega) 1 unit of GoTaq Flexi DNA Polymerase (Promega) to 4µl 5x Buffer (Promega) and adjusted the final volume to 20µl using H_2O. PCR conditions were as follows: Initial denaturation of 5min at 95°C, 35 cycles of 30 second denaturation at 95°C, 45 seconds annealing at 53°C, 2 min extension at 72°C followed by a final extension of 2 min at 72°C. PCR products were visualized in a 1 % agarose gel. As the degenerate primers created multiple bands, the largest band nearest to the expected insert size (approximately 1000bp) was excised and purified using the Qiagen Gel Extraction Kit. These were then cloned using the TOPO TA Cloning Kit (Invitrogen). Fifty colonies for each species were picked and lysed in 50 µl of H_2O (5 minutes at 95°C). Lysed colonies were then PCR amplified using in 1.1 x Reddymix PCR Master Mix (1.5 mM, Thermo Scientific) and 0.2mM of each M13 primer in a final volume of 15 µl. PCR products were purified for Sanger sequencing using Exonuclease I (Exo, Biolabs) and Thermosensitive Alkaline Phosphatase (TSAP, Promega).

Forward and reverse sequences for each clone were assembled into a single contig and subsequently cleaned using the chromaseq package in Mesquite (Maddison & Maddison 2014) prior to performing a blast search (Madden 2003) against the NCBI nucleotide and protein databases to confirm the identity of the amplified fragment.

Alignments were created using Muscle (Edgar 2004), and Corydoradinae specific primers (SH1/SH2) were designed using primer 3 (http://primer3.sourceforge.net/) (see table 1). These were then optimized using a gradient PCR and subsequently used for further PCR and TOPO TA cloning runs. In total, PCR products for eighteen species across the Corydoradinae phylogeny were amplified and cloned. Per species, between ten and thirty clones were randomly selected for Sanger sequencing with both forward and reverse primers. Amplification for these primers was as outlined above.

Table 1. Primers used to amplify *HoxA13a*. H15/16 is a degenerate primer pair after Yuan et al. (2010). SH1/2 were designed specifically for the Corydoradinae as part of this study.

Primer	**Forward**	**Reverse**
H15/16 (*HoxA13b*)	CTGGATTGACCCGGTSATGTT	TGRAACCAGATDGTSACYTGTCG
SH1/2	GATTGACCCGGTSATGTTC	CTGCGCTTGTCCTTGGTAAT

Data Analysis

Sequence Analysis

Forward and reverse sequences for each clone were assembled into a single contig and subsequently cleaned using the Chromaseq package in Mesquite (Maddison & Maddison 2014) and then aligned using Muscle as above. The Chromaseq package uses Phred (version 0.020425) and Phrap (version 0.990622) (Ewing et al. 1998; Ewing & Green 1998) which gives error probabilities for base calling. All indicated regions of uncertainty (below a quality score of 20) were checked visually. If areas of uncertainty stretched multiple bases at the front or the end of the read, they were trimmed. Reads with poor quality across the entire read were removed and excluded from further analysis. Sequences were dereplicated using Usearch (Edgar 2010) with the ‘derep_fulllength’ and were checked for the presence of chimeras using

UCHIME in the self-detection mode (Edgar et al. 2011). In order to distinguish between genuine alleles and sequencing errors, we calculated the probability of a given number of SNPs in an amplicon using a cumulative binomial distribution in Excel (Cummings et al. 2010) and used a Bonferroni corrected p-value cut-off to determine alleles. We retained sequences that were below the critical p-value if the sequence occurred more than 5 times or proportionally constituted more than 40% of sequences within the species block. As only one individual per species was cloned and sequenced, putative number of alleles was taken as a proxy for the minimum copy number within a species.

Phylogenetic Analysis

All identified allele-candidates were analysed in jModeltest version 2.1 (Darriba et al. 2012) to determine appropriate nucleotide substitution models. MrBayes version 3.2.2 (Ronquist et al. 2012) was run for 1,000,000 generations using the identified model (K86+G) and a sample frequency of 500. The equivalent model GTR-GAMMA in RAxML version 8.1.16 (Stamatakis 2014) was executed for 100000 rapid bootstrap inferences followed by a thorough ML search. Resulting trees were visualized in FigTree version 1.4.2 (Rambaut 2015).

Testing for signs of selection and functional divergence

In the absence of *HoxA13*a mRNA sequences from the Corydoradinae, we used the mRNA alignment from *Ictalurus punctatus* channel catfish to identify coding sequences in our alignment, indicating that the first ~600 bases of the alignment are coding. The coding region identified stretches beyond the *Ictalurus* alignment. Additionally, the NCBI ORF Frame finder (Tatusov & Tatusov 2015) was used to identify further potential coding regions by identifying potential start and stop codons within the sequence. Such application indicated that there is potentially another small coding region starting at ~800 bases until the end of the sequence.

We used Selecton (Stern et al. 2007) to calculate ratios of synonymous (d_S) and non-synonymous (d_N) substitutions across our coding regions. The analysis was conducted using both the first exon identified through the mRNA alignment as well as the second exon suggested by the ORF frame reader using the default M8 model (which allows for positive

selection) described in Yang et al. (2000). Selecton uses a Maximum Likelihood Approach to calculate codon specific d_N/d_S ratios (ω) and uses a Bayesian approach to calculate posterior probabilities.

We also downloaded *HoxA13a* orthologs for 18 different teleost species from Genbank (see table 2), carefully selecting only diploid fish species. We repeated the Selecton analysis for this data set in order to be able to compare selective pressures with pressures acting on the orthologous gene in species that have not undergone duplication. We aligned all other selected teleost species against the *Ictalurus punctatus HoxA13a* copy and trimmed sequences accordingly. A Kruskal-Wallis test was conducted to identify significant differences in ω ratios between the Corydoradinae and the outgroup-teleost species.

We used DIVERGE3.0 (Gu & Vander Velden 2002; Gu et al. 2013) to test for functional divergence between paralogous groups. DIVERGE is a phylogeny based software that can distinguishes between two types of functional divergence: Type I divergence is indicative of different evolutionary rates at sites of two duplicated genes. Type II divergence identified fixed changes in codon state, i.e. changes that lead to changes in its charge or lead to changes from hydrophobic to hydrophilic.

Table 2. Teleost species used in the Selecton analysis. Accession Codes from the NCBI and Ensemble databases.

Species name	Common Name	Superorder	Order	Accession Code
Ictalurus punctatus	Channel Catfish	Ostariophysi	Siluriformes	gi|222708662
Haplochromis burtoni	-	Acanthopterygii	Perciformes	gi|545786629
Oryzias latipes	Japanese Rice Fish	Acanthopterygii	Beloniformes	gi|429900369
Danio rerio	Zebrafish	Ostariophysi	Cypriniformes	gi|685508223
Astyanax mexicanus	Mexican Tetra/Blind Cave Fish	Ostariophysi	Characiformes	gi|597755542|
Apteronotus leptorhynchus	Brown Ghost KnifeFish	Ostariophysi	Gymnorynchiformes	gi|222708660
Cyprinus carpio	Common Carp	Ostariophysi	Cypriniformes	gi|685042158
Esox lucius	Northern Pike	Protacanthopterygii	Esociformes	gi|742193615
Poecilia reticulata	Guppy/Millionfish/Rainbowfish	Acanthopterygii	Cyprinodontiformes	gi|658873801
Megalobrama amblycephala	Wuchang Bream	Ostariophysi	Cypriniformes	gi|123204431
Chanos chanos	Milkfish	Ostariophysi	Gonorynchiformes	gi|222708658
Gasterosteus aculeatus	Three-spined stickleback	Acanthopterygii	Gasterosteiformes	ENSGACT00000009477
Gadus morhua	Atlantic Cod	Paracanthopterygii	Gadiformes	ENSGMOT00000015205
Tetradon nigroviridis	Green Spotted Pufferfish	Acanthopterygii	Tetraodontiformes	ENSTNIT00000012258
Xiphophorus maculatus	Southern Platyfish/Moonfish	Acanthopterygii	Cyprinodontiformes	ENSXMAT00000001052
Poecilia formosa	Amazon Molly	Acanthopterygii	Cyprinodontiformes	ENSPFOT00000002357
Takifugu rubripes	Japanese Puffer/Tiger puffer	Acanthopterygii	Tetraodontiformes	ENSTRUT00000041073
Oreochromis niloticus	Nile Tilapia	Acanthopterygii	Perciformes	ENSONIT00000005797

2.3 Results

Forward and reverse sequencing of 402 products for 18 different species resulted in 55 consensus sequences/potential alleles after dereplication and cleaning (table 3).

Sequence Identification, Diversity and Putative Alleles

All sequences were identified as *HoxA13*a using the online NCBI discontinuous nucleotide megablast program (Madden 2003) with the top hits including *Ictalurus punctatus* (the channel catfish¸ and *Danio rerio* (Zebrafish). Both *I. punctatus HoxA13a* and *HoxA13b* sequences were downloaded from the NCBI database and used as outgroup sequences for phylogenetic analysis.

Table 3. Species, the number of sequences obtained that passed quality filters as well as the putative alleles identified. C-values are also listed for all species.

Lineage	Species	Sequences	Potential Alleles	C-value
1	*C. semiaquilus*	14	2	0.51
2	*Aspidoras C110*	23	1	0.76
3	*S. kronei*	9	5	0.8
3	*S. macropterus*	22	4	0.82
4	*C. hastatus*	24	5	2.2
5	*C. nijsseni*	21	1	2.02
6	*C. paleatus*	18	5	1.62
6	*C. tukano*	23	1	2.48
7	*C. aeneus* (Green Laser)	24	1	1.84
7	*C. rabauti*	22	2	2.05
8	*C. imitator*	24	2	2.3
8	*C. leopardus*	22	2	1.96
9	*C. adolfoi*	30	5	3.77
9	*C. armatus*	30	6	4.3
9	*C. metae*	27	3	4.15
9	*C. araguaiaensis*	21	5	4.36
9	*C. arcuatus*	22	2	2.28
9	*C. axelrodi*	26	3	3.2
	Total	**402**	**55**	

The number of potential alleles identified ranges from only 1 allele in *A.* C110, *C. tukano* and *C. aeneus* to a maximum of 6 alleles (found in *C. armatus)*. In total, nine samples belonging to 9 species contain more than two alleles. Surprisingly, both lineage 3 species *S. kronei* and *S. macropterus* contained 5/4 alleles; more alleles than we initially expected as genome size is only marginally larger than that of *C. semiaquilus*. *C. hastatus* with a genome size almost three times the size of that of the *Scleromystax* species also has 5 alleles, as does *C. paleatus* with a genome size of 1.62. All other individuals containing multiple alleles belong to most derived lineage 9. Despite large genome sizes, only 1 allele was found for *C. aeneus* (Green Laser) and all remaining individuals. We cannot exclude the possibility of failure of amplification of further copies, which could lead to the underestimation of sequence diversity. However, we can use individuals in which we found more than 2 alleles to infer duplication events across the phylogeny, based on the assumption that *Hox* genes do not undergo tandem duplications.

Phylogenetic Analysis and Identification of Paralog Genes or Paralogous Groups

Both MrBayes and RAxML recovered the same topology for the allele tree and exhibiting nine supported monophyletic clades shown in Figure 3. Six out of these clades contain alleles belonging to a mixture of different species belonging to different lineages (see table 4). We consider these clades distinct paralogs, due to their high degree of divergence and as we would expect sequences to cluster by gene copy rather than by species. Alternatively, the similarity could also be explained by very recent hybridization events among species, but no hybrids have been reported within the Corydoradinae. In view of the age of different lineages, as well as the number of species involved and their different karyotypes, we consider this less likely. Thus we have identified six different potential *HoxA13a* paralogs, containing between two and six different species belonging to two to five different lineages. The three remaining clades belong to the same species or in case of clade 8 to the same lineage. Furthermore, we have divided these clades into four groups according to the last common node they share, which should represent a duplication event (Figure 3). For instance, clades 7, 8 and 9 share one node and form group 4 (table 4**).**

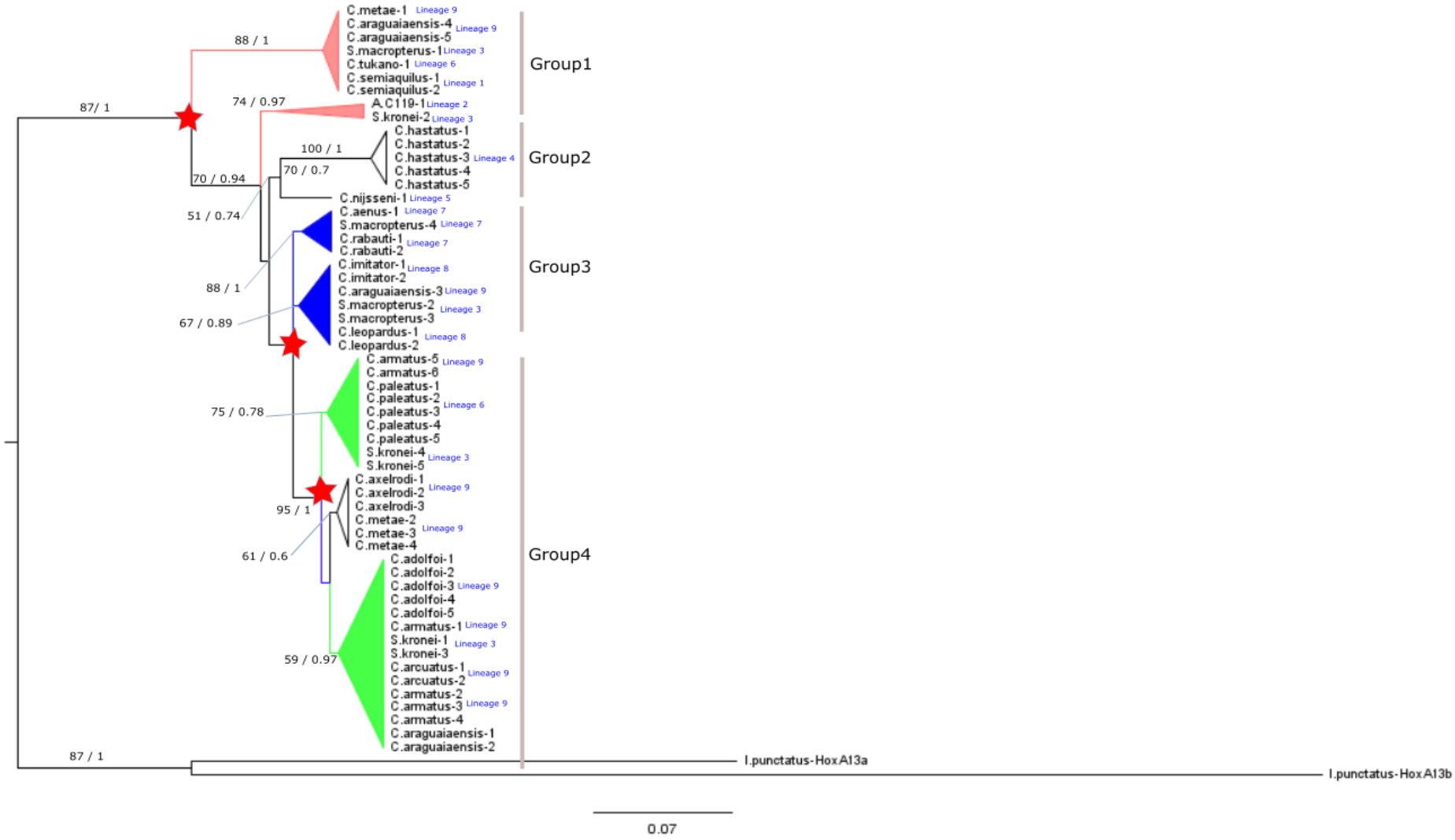


Figure 3. Topology recovered from the phylogenetic analysis. Values near nodes represent the Bootstrap support in RAxML and the posterior probabilities in MrBayes (Bootstrap / Posterior Probability). Clades/putative paralogs that contain a mixture of different species from different lineages are displayed in colour. Paralogs are grouped according to their potential point of origin (marked with a star).

Note that we grouped clades 1 and 2 together, even though they do not strictly share a node. However, we consider clades 3 and 4 clearly distinct, and containing only two sequences we did not think it appropriate for clade 2 to be a distinct group. We therefore added it to the most closely related clade 1. Group 2 is unique in that it contains species from lineage 4 and 5 that do not cluster with any other species. The position of group 2 is also the least well supported in both the RAxML and MrBayes analysis. All other groups represent a mix of different species and lineages. Species from lineages 3, 6 and 9 are found in all three groups, whereas species from lineages 1, 2, 7 and 8 are only found in one group overall. Group 1 contains the largest number of species and lineages and is also the most basal group, resembling the orthologous *HoxA13a* in *Ictalurus punctatus* the most. Group 2 exclusively contains the lineage 7 and 8 species, whereas group 3 is heavily dominated by lineage 9 species.

The fact that not all paralogous clades or even groups contain a member species of each lineage could be the result of a failure of amplification (null alleles) or potentially the loss of certain clusters in certain lineages or species. The basal lineage 1 species *C. semiaquilus* possesses two alleles, both of which are found in the basal clade 1. *A. C119* (lineage 2) appears in the separate Clade2, but in no other groups. Both *Scleromystax sp. (*lineage 3) are present in all clades and groups with the exception of the lineage 4/5 group. Clade 1 and 2 appear quite diverged, which could indicate that a duplication event preceded the split. Particularly as we would expect *S. kronei* and *S. macropterus* to cluster within the same group otherwise. As *A. C119* is not found in any more derived clades, it is a possibility that there have been one or several additional duplications after the lineage 2-lineage3 split.

While we cannot exclude null alleles or lineage specific cluster loss, duplications must have occurred prior to the lineage 3-lineage 4 split, as lineage 3 is indeed represented in all major clusters. All *C. hastatus* alleles cluster together and not with any other species, which could indicate a lineage or species-specific duplication event. As a lineage 4 species, it should share the above-mentioned duplications but either lost evidence of these paralogs or these failed to amplify. Lineage 9 specific duplications also appear evident in group 4, with *C. axelrodi, C. adolfoi* and *C. armatus* all possessing multiple alleles clustering together within this group. *C. paleatus* (lineage 6) shows a similar signature with 5 alleles in group 4.

Selection Analysis and Functional Divergence

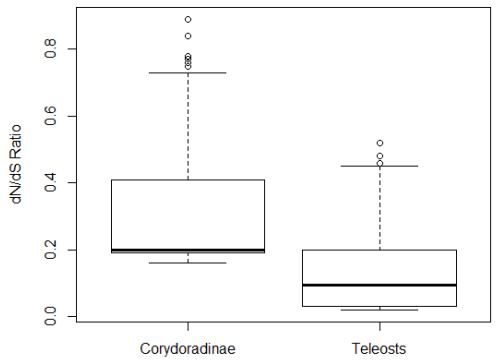
The Selecton analysis did not detect any signs of positive selection within any of the coding regions, but shows that a large number of amino acids appear under heavy purifying selection (figure 4a)**.** We compared this to *HoxA13a* orthologs in other teleosts that have not undergone additional rounds of WGD (figure 4b**)**. We tested for differences in the selection strength between the two groups by comparing the obtained Ka/Ks ratios of the Corydoradinae duplicated copies with those in other teleost species. We detected a significant difference that indicates a significant shift (Kruskal-Wallis, chi-squared =77.3662, df=1, p-value < 2.2e-16) in the selection strength as displayed in figure 5**.** Thus, while the Corydoradinae copies are not subject to positive selection, they are under significantly less strong purifying selection than their non-duplicated counterparts, and thus possibly evolving more quickly than the non-duplicated orthologous sequences in other teleosts.

We used DIVERGE to test whether the relaxed selection could have resulted in functional differences by comparing the four groups previously described (table 5). After Bonferroni correction, there were no significant changes in the Type II analysis. However, group 2 (containing lineages 4 and 5) appears to be significantly different from all other groups in the Type I analysis, indicating a significant shift in amino acid changes in this group. While there certainly also appears to be a strong signal of Type I divergence between all other groups, the significance does not hold after Bonferroni correction.

Table 4. All clades determined in the phylogenetic analysis, as well as their *Hox* groups, species composition and the mtDNA lineages they belong to.

Paralog Clades	Group	Sequences	Species	Lineages Present	Number of Lineages
Clade 1	1	*Corydoras metae 1, Corydoras araguaiaensis 4, Corydoras araguaiaensis 5, Scleromystax macropterus 1, Corydoras tukano 1, Corydoras semiaquilus 1, Corydoras semiaquilus 2*	5	1, 3, 6, 9	4
Clade 2	1	*Aspidoras C119, Scleromystax kronei 2*	2	2, 3	2
Clade 3	2	*Corydoras hastatus-1, Corydoras hastatus-2, Corydoras hastatus-3, Corydoras hastatus-4,Corydoras hastatus-5*	1	4	1
Clade 4	2	*Corydoras nijsseni*	1	5	1
Clade 5	3	*Scleromystax macropterus 2, Scleromystax macropterus 3, Corydoras imitator 1,Corydoras imitator 2, Corydoras araguaiaensis 3, Corydoras leopardus 1, Corydoras leopardus 2*	4	3, 8, 9	3
Clade 6	3	*Corydoras aeneus 1, Corydoras rabauti 1, Corydoras rabauti 2, Scleromystax macropterus 4*	3	3, 7	2
Clade 7	4	*Corydoras armatus 5, Corydoras armatus 6, Scleromystax kronei 4,Scerlomystax kronei 5, Corydoras paleatus 4, Corydoras paleatus 1, Corydoras paleatus 2, Corydoras paleatus 3, Corydoras paleatus 5*	3	3, 6, 9	3
Clade 8	4	*Corydoras metae 4, Corydoras metae 3, Corydoras metae 2, Corydoras axelrodi 1, Corydoras axelrodi 2, Corydoras axelrodi 3*	2	9	1
Clade 9	4	*Corydoras araguaiaensis 1, Corydoras araguaiaensis2, Corydoras armatus 2, Corydoras armatus 3, Corydoras armatus 4, Corydoras arcuatus 1, Corydoras arcuatus 2, Corydoras armatus 1, Scleromystax kronei 1, Scleromystax kronei 3, Corydoras adolfoi 1, Corydoras adolfoi 2, Corydoras adolfoi 3, Corydoras adolfoi 5, Corydoras adolfoi 4*	5	3, 9	2

Figure 4. Selecton analysis with each amino acid position being colour coded according to the calculated selection strength based on ω (d_N/d_S ratio). a) Amino acid alignments resulting from the Corydoradinae based on 56 sequences containing all identified coding regions. b) Amino acid alignment based on 18 teleost species with an alignment trimmed based on *Ictalurus punctatus.*



.

Figure 5. d_N/d_S ratios calculated for all sequences of both the Corydoradinae and the Teleostei.

Table 5. Functional divergence analysis conducted in DIVERGE. P-values that are significant after Bonferroni correction are highlighted in bold. *MFE Theta* =Estimate of θI by the model-free method. *MFE SE* = Standard error of the θI estimated by the MFE. *MFE r_X* = Observed coefficient of correlation between two gene clusters. *MFE z-score* = z-score for model-free estimate of the after Fisher's transformation. *P-value* = calculated using the z-score in a two-tailed z-score test. *N*= Number of sites with no change between two clusters. *C* = Number of sites with conserved change between two clusters. *R* =Number of sites with radical change between two clusters. *Alpha ML* = Maximum likelihood estimate of α. *Theta* = Estimate of θII by simplified maximum-likelihood method. *Theta SE* = Standard error of θII. Descriptions of parameters were taken from the Manual (Gu 2013).

Type I Functional Divergence Analysis						
	Group 1/Group 2	Group 1/Group 3	Group 1/Group 4	Group 2/Group 3	Group 2/Group 4	Group 3/Group 4
MFE Theta	0.500811	1.210499	0.622422	1.268689	0.946565	1.203017
MFE se	0.189664	0.602535	0.26614	0.424035	0.18933	0.582384
MFE r X	0.179235	-0.024566	0.098846	-0.044495	0.01985	-0.024512
MFE r max	0.359053	0.116702	0.261789	0.165601	0.371478	0.120741
MFE z score	-2.765873	-2.015461	-2.399927	-3.008318	-5.262727	-2.07292
P-value	**0.005692**	0.043905	0.0164	**0.002627**	< **0.00001**	0.037182
Type II Functional Divergence Analysis						
	Group 1/Group 2	Group 1/Group 3	Group 1/Group 4	Group 2/Group 3	Group 2/Group 4	Group 3/Group 4
N	188	193	193	193	195	201
C	9	4	6	8	8	2
R	8	8	6	4	2	2
Alpha ML	0.133903	0.653439	0.345533	0.138032	0.09553	0.719395
Theta-II	0.013897	-0.004912	0.000592	0.066713	0.041352	0.018609
Theta SE	0.031787	0.04039	0.034901	0.032592	0.028261	0.034793
P-value	0.662039	0.903691	0.986516	0.040668	0.143413	0.592788

2.4 Discussion

Here, we identify several putative paralogs of *HoxA13a* paralogs, indicating that several rounds of WGD events took place in the Corydoradinae. Despite having undergone several rounds of duplication, we do not find any evidence for positive selection and instead find that all paralogs remain under strong purifying selection. However, when compared to other teleosts who have not undergone recent additional rounds of WGD, the purifying selection is significantly less strong, and our data indicates some degree of functional divergence not found in previous studies.

Ohno (1970) himself noted the fourfold increase and the variability in genome size in Corydoradinae catfish in comparison with the Channel Catfish *Ictalurus punctatus* in his famous book *Evolution by Gene Duplication* and speculated that -due to the similarity in karyotype- the genome size expansions must be the result of tandem gene duplications, as opposed to WGD events. However, recent phylogenetic work by Alexandrou et al. (2011) indicates stepwise shifts in genome size between lineages which could also be explained by WGD events, if followed by subsequent chromosomal rearrangements and rediploidization.

Hox genes are very strictly conserved in their order within clusters, a property that has been used to infer WGD events in vertebrates (reviewed in Glasauer & Neuhauss 2014). More recently Crow et al.(2012) used *Hox* genes to demonstrate that the American Paddlefish had undergone an additional round of WGD. To our knowledge, no tandem duplications of *Hox* genes have ever been reported within vertebrates, and the possibility of tandem duplications in the previously analysed and closely related zebrafish (*Danio rerio*) *Hox clusters* has been excluded (Amores et al. 1998). Garcia-Fernàndez & Holland (1994) suggested that the strict conservation of gene order is likely related to precise spatio-temporal expression during embryonic body development. Tandem duplication within a cluster would thus potentially disrupt or significantly alter the function of the cluster in early embryonic development. Thus, we believe *Hox* genes could also serve as a reliable indicator of WGD events within the Corydoradinae.

Alternatively, the Corydoradinae could serve as the first example of tandem duplications of *Hox* genes within vertebrates, or indeed multiple independent hybridization events between highly divergent species with different genome sizes and karyotypes. For hybridisation to explain our data, at least 11 out of 18 species would have had to be involved

in a recent hybridization event**.** In the context of previous work on teleosts and considering the divergence between Corydoradinae species, we believe that the occurrence of multiple WGD events is the most parsimonious explanation for our data.

Up to three WGD events or perhaps several triploidy events are likely somewhere between the split of lineages 1-3. We believe it is most likely to have occurred after splitting from lineage 1, though we cannot entirely exclude the possibility that *C. semiaquilus* has not simply lost all other paralog-copies. In fact, either a failure to amplify all paralog copies for all species (plausible as we only used one set of primers), or a large amount of species or lineage specific gene loss is apparent in our data. Multiple allele copies of *C. paleatus, C. adolfoi, C. hastatus* and *C. metae* are all present within the same clade or group, indicating that these species have undergone an additional lineage specific duplication and these multi-copy alleles have yet to diverge into distinct paralogs or be lost. However, we would expect to see these multi-copy alleles in all groups in which species are present, which we do not. Furthermore, lineage 7 and 8 species appear to only be present in group 2. Lineage or species specific gene loss would not be surprising, as a large scale loss of *Hox* genes has also occurred shortly after the FSGD event and may still be an ongoing process (Duboule 2007; Prohaska & Stadler 2004; Kuraku & Meyer 2009). While teleosts certainly contain more *Hox* clusters than other vertebrates, Duboule (2007) pointed out that the gene number they contain is fairly comparable: For instance, teleosts possess roughly 48 genes in 13 clusters and mice possess 38 genes in 4 clusters. The *Hox* gene content has been shown to vary considerably between different teleost species (Kuraku & Meyer 2009).

The most likely fate of duplicated genes is subsequent loss (Lynch & Conery 2000) and only a very small proportion of genes are expected to be retained. If they are retained, they could be subject to sub-functionalization (i.e. both copies take over part of the ancestral gene's function) or neo-functionalization (acquiring a new function after being freed from purifying selection pressure). Since Ohno (1970) first postulated his hypothesis, it has been a common notion that gen(om)e duplication is a major mechanism creating novel gene functions. However, initial genome and taxon- wide scans show that most gene duplicates remain under purifying selection, though this selection pressure is generally weaker than in non-duplicated orthologs (Lynch & Conery 2000; Kondrashov et al. 2002; Prohaska & Stadler 2004). Even after multiple duplications of the *HoxA13a* gene, the paralogs still appear to be under purifying selection, with no evidence of positive selection being found in any of our data. Similarly to results reported in literature (Prohaska & Stadler 2004; Kondrashov et

al. 2002), we find that purifying selection pressure is indeed significantly weaker than in its non-duplicated teleost counterparts.

Positive selection was not detected in either *HoxA13a* in zebrafish, or in any *Hox* clusters in Atlantic salmon (*Salmo trutta)* (Crow et al. 2009; Mungpakdee et al. 2008)**.** *HoxA13a* appears to be evolving much faster than its paralog *HoxA13b* (Crow et al. 2009), and several studies have demonstrated that *HoxAa* clusters appear to show significantly higher K_S values when compared to other clusters (Wagner et al. 2005; Mungpakdee et al. 2008). Thus, we cannot exclude the possibility that positive selection may merely be masked by a high synonymous substitution rates. However, even in the absence of positive selection, functional divergence can occur. Such diversity could arise either through neutral processes, or because relaxed purifying selection may lead to a latent build-up of substitutions eventually creating novel functions.

When looking for signs of functional divergence, we did not find any significant Type II changes, i.e. site-specific shifts in amino acid properties. Instead, we find highly significant Type I functional divergence between group 2 (*C. hastatus* and *C. nijsseni*) and all other groups. It is interesting that group 2 has diverged significantly from all other lineages, and is perhaps related to an additional round of WGD in this group. While lineage 9 also appears to have undergone an additional WGD event, the duplication in lineage 5 is much older (based on the age of the lineage), and would have thus had more time to diverge. However, such an assertion is only speculative, and warrants further investigation in future studies. While not significant after the Bonferroni correction, there is still a considerable signal of Type I functional divergence between the three remaining groups, indicating that *HoxA13a* paralogs in the Corydoradinae are functionally diverged, or in the process of becoming so. Crow et al. (2009) did not detect any Type I divergence in the *HoxA13a* Ostariophysian clade (which included the Channel Catfish). Potentially, therefore, the additional rounds of WGD have allowed for functional divergence within the Corydoradinae.

Conclusions

Next Generation Sequencing (NGS) technologies are revolutionizing the fields of molecular ecology and evolution (Mardis 2008; Stapley et al. 2010; Mardis 2011; Ekblom & Galindo 2011; van Dijk et al. 2014), and open the door to study polyploid or paleopolyploid

organisms in unprecedented depth. However, in order to utilize such technology efficiently, it is essential to be aware of genome sizes and ideally genome copies, to design experiments and scale sequencing efforts appropriately. Particularly for species with a polyploid history, sufficient sequence coverage becomes crucial in order to distinguish between paralogous copies (e.g. Hohenlohe et al. 2012)**.** Prior to the availability of genome data and NGS technology, *Hox* genes served as a powerful initial evidence for WGD events in vertebrate evolution. Now that their status as a marker for WGD is generally accepted (Crow et al. 2012; Glasauer & Neuhauss 2014)**,** they serve as an insightful marker to detect WGD events and help design further research questions in species that lack prior genomic information.

In summary, we show that

1) *HoxA13a* serves as an informative marker for detecting WGD events in the Corydoradinae.
2) Multiple WGD events appear to have occurred within the family, some specific to certain lineages, and some shared between lineages.
3) *HoxA13a*-paralogs are under significantly less strong purifying selection than their non-duplicated teleost orthologs.
4) Relaxed purification appears to have led to Type I functional divergence between Corydoradinae paralogs, a pattern not seen in other teleost orthologs examined to date.

2.5 Supporting Information

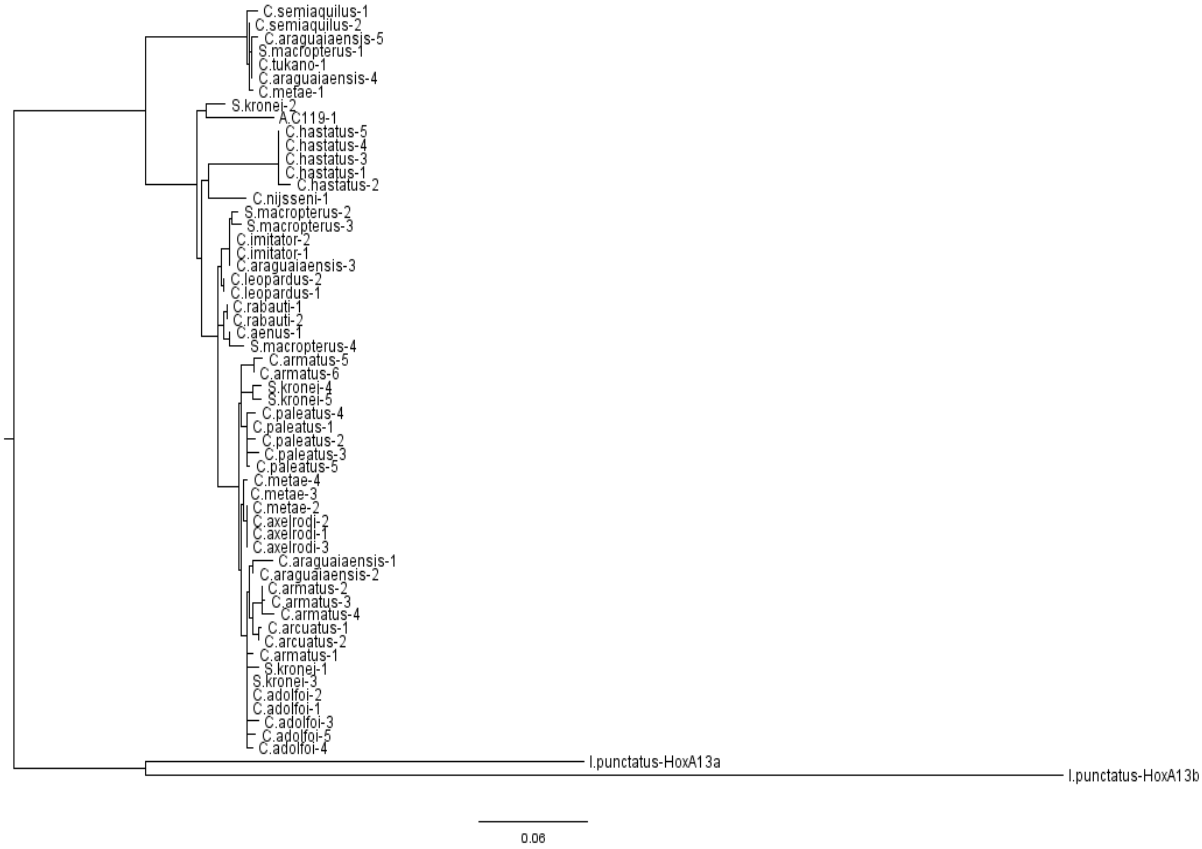


Figure 6. Topology afterRaxMLanalysis. Bootstrap values not shown for clarity.

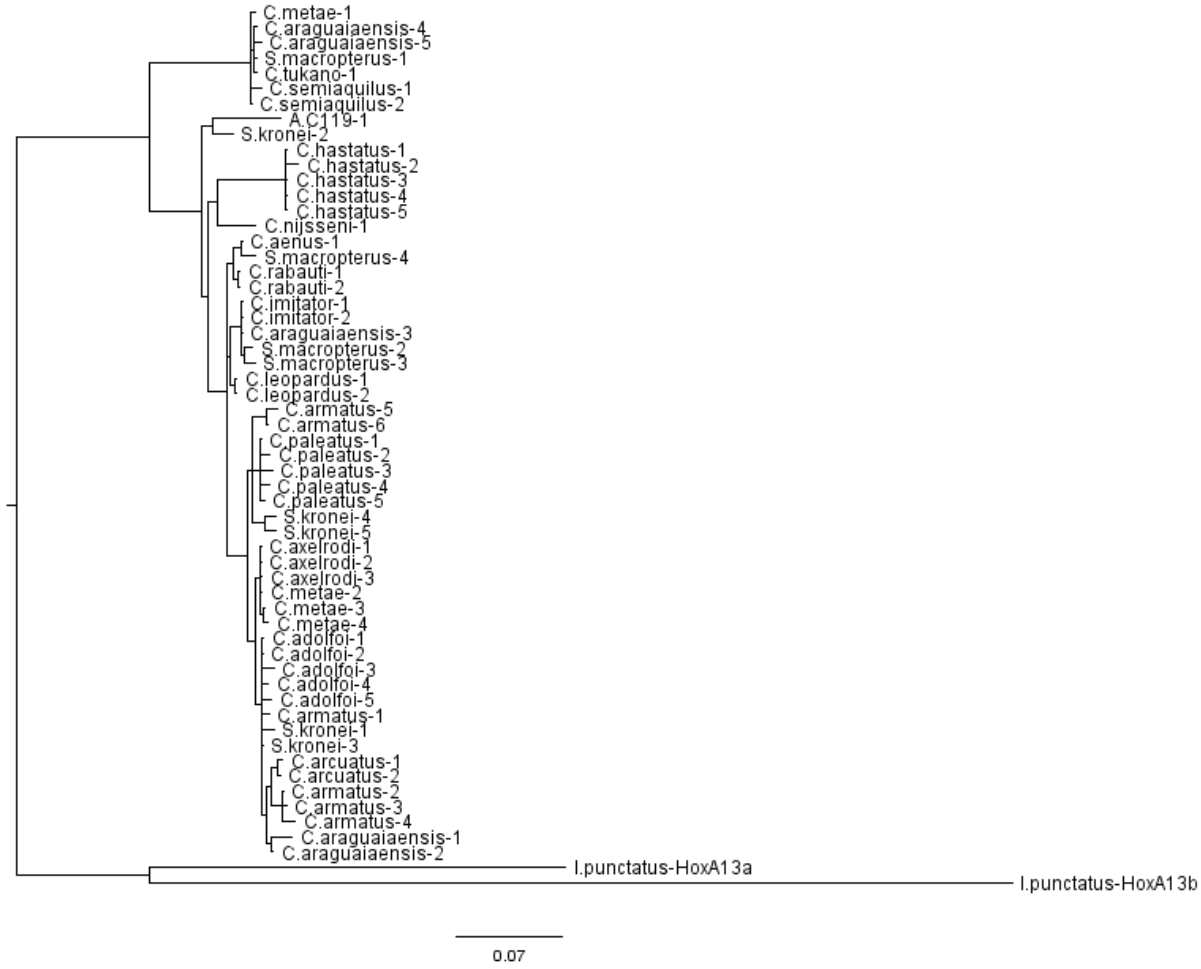


Figure 7. Topology after MrBayes analysis. Node posterior probabilities are not shown for clarity.

Chapter 3 - Two rounds of whole Genome Duplication in the Corydoradinae Catfish detected using RAD Sequencing

3.1 Introduction

Whole Genome Duplication (WGD) or polyploidization is the addition of a complete set chromosomes following hybridization (allopolyploidy) or through doubling of the genome within a species (autopolyploidy). WGD is one of the most severe genetic mutations that can occur in an organism with potentially severe consequences such as genome instability, large scale genomic rearrangements and deletions, changes in gene regulation and TE activation (Otto 2007). While originally perceived as an evolutionary dead-end (Stebbins 1950; Stebbins 1971), it is now clear that ancient WGD events occurred in the ancestors of the majority of plants as well as all vertebrates (Otto & Whitton 2000; Mable 2003; Gregory & Mable 2005; Mable et al. 2011). Two rounds of WGD (often referred to as 1R and 2R) appear to have preceded the radiation of vertebrates (Ohno et al. 1970; Putnam et al. 2008; Furlong & Holland 2002), with a third WGD event (termed 3R, FSGD or TS-WGD) occurring at the base of the teleostei radiation (Amores et al. 1998; Meyer & Van de Peer 2005; Crow et al. 2006; Kasahara et al. 2007; Glasauer & Neuhauss 2014). These ancient WGDs can be difficult to detect in the absence of full genome sequences, as polyploid species return to a functionally diploid state over time - a process known as re-diploidization (Wolfe 2001). In addition to well established ancient WGDs, many more recent events have been documented in extant plants, amphibians and fish species which are considered functionally polyploid (reviewed in Mable *et al.* 2011; Glasauer & Neuhauss 2014).

Susumu Ohno (1970) was one of the first prominent advocators of the hypothesis that gene and genome duplications create novel material which could result in the development of new genes and adaptive traits. While some immediate genetic consequences of WGD events such as extensive genomic rearrangements, chromosome loss and changes in gene expression are becoming clearer (see Otto [2007] for a review), the long-term evolutionary consequences and impact of polyploidy on the diversification and the adaptive potential of a species remain a topic of much debate (Santini et al. 2009; Mayrose et al. 2011; Soltis et al. 2014; Zhan et al.

2014; Glasauer & Neuhauss 2014). While WGDs appear to be more frequent in the evolutionary history of both plants and vertebrates than previously accepted, time lags between such events and subsequent radiations cast doubt on the direct influence of WGDs as a driver of evolutionary novelty. Instead, it has been suggested that WGDs could provide the necessary genomic environment for later radiation and adaptive evolution to occur (Glasauer & Neuhauss 2014). To better understand the evolutionary consequences of WGDs, as well as the ability of genomes to return to a functionally diploid-like state, a system that comprises independent WGD events, or several lineages evolving after a common WGD event at evolutionary distinct times, could help elucidate different mechanisms governing the long term consequences of WGD in a statistically robust manner.

Next Generation Sequencing (NGS) has provided evidence for the broader occurrence of WGD and polyploidy across eukaryotes, and enabled evolutionary biologists to study the retention of duplicate genes and genome dynamics in unprecedented detail (Blanc & Wolfe 2004; Jaillon et al. 2004; Renny-Byfield & Wendel 2014; Soltis et al. 2009). Many techniques have been developed that allow for the utilization of genomic data even in the absence of reference genomes, enabling the in-depth study of non-model organisms (reviewed in Davey et al. 2011). One such technique is Restriction Site Associated DNA (RAD) Sequencing (Baird et al. 2008; Etter, Preston, et al. 2011; Etter, Bassham, et al. 2011). RAD was initially developed for SNP discovery and genotyping in non-model organisms, but has since been successfully applied to phylogenomics, phylogeography, genome scaffolding, population genetics and linkage mapping (Davey et al. 2011), detecting introgression between species (Eaton & Ree 2013) and the mapping and studying of evolutionary adaptive traits (Hohenlohe et al. 2010; Recknagel et al. 2013).

One group frequently suggested as being polyploid based on chromosome counts and genome size variation are the Corydoradinae catfishes (Oliveira et al. 1992; Otto & Whitton 2000; Mable et al. 2011). The Corydoradinae catfish are the most diverse group of the order Siluriformes, with more than 170 described species (Eschmeyer 2013) and display large variation in genome size with C-values ranging from 0.5 to 4.8 pg of DNA (Alexandrou 2011). Such observations have led to the hypothesis that several WGDs could have occurred within the Corydoradinae. Phylogenetic analysis places the Corydoradinae into 9 distinct mtDNA lineages, where higher C-values are found in the more derived and younger lineages (see figure 2, Chapter 1; Alexandrou 2011). Identifying the number and timing of potential

WGDs in the evolutionary history of the Corydoradinae is an important step in understanding drivers of their diversification. The aim was to use RAD sequencing to identify potential WGD events in the Corydoradinae subfamily, enabling inferences on evolutionary diversification in relation to genome size. Specifically, the objectives were

1) to find evidence for WGD events across the genome (to complement the *Hox* gene-based findings of chapter 2),
2) to identify whether species that have undergone WGD events remain functionally polyploid,
3) to create a phylogeny using the RAD data which could help identify potential hybridization events.

3.2 Methods

Ploidy-Analysis – Nomenclature and Rationale

A common characteristic of polyploidy is the change from a disomic to a tetrasomic inheritance and segregation pattern, which results from homeologous chromosomes pairing at meiosis (Stebbins 1971). For instance in a recent tetraploid, a given locus may contain four segregating alleles. However, in an older polyploidy where homeologous pairs have diverged, or in an allopolyploid where chromosome sets from parental species are sufficiently diverged to begin with, preferential chromosome pairing of homologs occurs, leading to bivalent formation and disomic inheritance (Stebbins 1971). In this case, a given locus would behave like a diploid locus with two segregating alleles. It is furthermore possible for a mixture of both tetrasomic and disomic inheritance pattern to occur in the same organism, such as for instance in Atlantic salmon (*Salmo salar)* (Allendorf et al. 2015).

In addition to duplication through WGD events, genomic regions and genes can be duplicated by other means such as for instance through tandem duplication events (Taylor & Raes 2005). Duplicated Regions or loci are referred to as paralogs. Duplicated regions through WGD events are referred to as ohnologs (named after Susumu Ohno to honour his work on WGD in vertebrates) (Wolfe 2000) and are as such a special type of paralogs.

Based on short read RAD data and in the absence of a reference or detailed ploidy information, it is not possible to distinguish between ohnologs and paralogs. Thus, any contigs with more than two haplotypes are referred to as potential paralogs. As an assembled contig may contain several distinct coding segments, or indeed be entirely non-coding, the term allele may be inappropriate and thus, consecutive stretches of DNA that are inherited together, are referred to as haplotypes in this chapter.

Samples and DNA Extraction

To investigate whether WGDs have occurred in the Corydoradinae and if they have, how many, samples were chosen from across each of the nine previously identified mtDNA lineages (figure 2 in chapter 1; Alexandrou 2011). These samples covered C-values ranging from 0.65 pg to 4.36 pg allowing species that range considerably in genome size to be investigated for signals of WGD. For lineages 1 to 8, one species per lineage was selected for RAD sequencing. For lineage 9, two species were selected with the highest genome sizes. *Megalechis sp.* was chosen as the outgroup and belongs to the same family (Callichthyidae) but not the same subfamily (Callichthyinae vs. Corydoradinae). For all species, two individuals were sampled and every individual sample was replicated once except for the outgroup *Megalechis*, where one replicated individual was used. Libraries were created in two library sets (set 1 & set 2), with 5 species and 6 species respectively (table 6). For RAD sequencing, it is crucial to use high quality DNA with a high molecular weight, to avoid problems with the restriction enzyme digestion and ensure the overall success of the protocol (Etter et al. 2011). Thus, several species for each lineage were tested and those with the best DNA quality were selected for this study. Unfortunately, this meant that the same species used in the previous chapter 2 for lineages 2, 4, 5 and 6 could not be used.

Each individual sample in set 1 was uniquely barcoded using a 5bp or 7bp sequence as part of the P1 adapter. As library set 2 consisted of 24 samples, we uniquely matched P1 with a combination of two different P2 adapters. Samples were pooled by individual species during library preparation, allowing the adjustment of the pooling strategy of finished libraries to compensate for variation in genome size.

Samples were obtained from archived wild collections stored in 95% Ethanol made by Martin Taylor and Claudio Oliveira between 2005 and 2013, or in a few cases from wild-caught aquarium trade sourced specimens. DNA was extracted using the Qiagen DNA Blood & Tissue Extraction Kit. All samples were treated with RNase and were selected for high quality and high molecular weight using 2 % agarose gels.

RAD Library Construction

RAD libraries were prepared as described in Etter et al. (2011) with minor modifications: 600ng of DNA per sample were digested in separate reactions using 0.21µl SbfI restriction enzyme (NEB, 20 units/µl), 12 µl DNA, 5ul Buffer X 10, 14.79 µl H_2O for 45 minutes at 37°C. To each sample, a specific P1 –Adapter containing a unique barcode was then ligated (0.6µl NEB2 Buffer, 0.36µl 100mM rATP, 1.44 µl H_2O, 1.5µl P1 Adaptor, 0.3µl 2MU/ml T4Ligase). Each individual sample in library set 1 was uniquely barcoded using a 5bp or 7bp sequence as part of the P1 adapter. As library set 2 consisted of 24 samples, we uniquely matched P1 with a combination of two different P2 adapters. Samples were incubated for 60 minutes at room temperature, followed by a 20 minute heat inactivation step at 65°C and then pooled by species. Samples were then sheared using a Covaris sonicator using the following settings: Duty cycle 10%; Intensity 5; Cycles/Burst 200; Mode: Freq sweeping; Duration 105 sec. A Qiagen Mini-elute Kit was used to concentrate samples into ~30µl of Elution Buffer (EB), which were run out in a gel to perform a size selection step. DNA between 200-550 bases was isolated and extracted using the Qiagen Mini-elute Gel Extraction kit. The sheared ends of the DNA fragments were then repaired using the NEB Quick Blunting Kit (2.5µl Quick Blunting Buffer, 2.5 µl 1mM dNTP mix, 1µl Blunt Enzyme Mix; incubated at room temperature for 40 minutes). All samples were subsequently purified using the Qiaquick Mini-elute kit and eluted into 43µl of EB and 5 µl 10X NEB Buffer 2. In a 45 minute incubation step at 37°C, dA overhangs were added to the DNA fragments (1µl 10mM dATP, 3µl Klenow exo) to prepare samples for the P2 ligation. A subsequent Qiaquick cleanup was performed and samples were eluted into 45µL EB Buffer and 5µL 10X Buffer 2 to which 1µL of 10µM P2-Adapter was added, as well as 0.5 µL 100mM rATP and 0.5µL concentrated T4 DNA Ligase. Reactions were incubated at room temperature for 45 minutes. An additional clean-up reaction was performed prior to high-fidelity PCR amplification optimization. Once test PCRs produced satisfactory bands, a final PCR step was performed (8.1µl H_20, 2x Phusion High Fidelity Mastermix 12.5µl, 0.4µl Primers). For each pooled

species-library, 16 x 12.5µl amplifications were performed using the following settings: 98°C 30s then 98/66/72 °C for 10s/30s/30s respectively for 14 cycles then 5 min at 72 °C. All PCRs for one library were combined and concentrated using the MiniElute PCR clean-up kit and another size selection step was performed using gel electrophoresis (300-600 bases). In the case of the RAD library set 2, a further gel extraction was necessary prior to PCR amplification due to a leftover adapter band at approximately 70 bps.

Sequencing

Each set of libraries was sequenced on one lane of the 150bp paired-end Illumina HiSeq2000 at The Genome Analysis Centre (TGAC), Norwich. In order to achieve even coverage for species of different genome sizes, species libraries were combined for sequencing in proportion to their C-values (table 6).

Table 6. List of species in each sequencing run. Libraries for each species were combined at different ratios to compensate for differences in genome size.

Sequencing Run	Lineage	Species	C-value	Sequencing Ratio
1	Outgroup	*Megalechis sp.*	1.58	1
1	1	*C. fowleri*	0.65	1
1	8	*C. imitator*	2.3	4
1	9	*C. metae*	4.15	8
1	9	*C. araguaiaensis*	4.36	8
2	2	*A. poecilius*	0.76	1
2	3	*S. kronei*	0.8	1
2	4	*C. pygmaeus*	2.68	3
2	5	*C. elegans*	2.24	3
2	6	*C. nattereri*	1.79	2
2	7	*C. aeneus*	1.84	2

Data Analysis

Data Cleaning and De-multiplexing

Raw sequences were cleaned using Trimmomatic (Bolger et al. 2014) with the following settings: PALINODROME mode to detect and remove adapter contamination, SLIDINGWINDOW was set to 4 and trimmed the read if the quality dropped below a set threshold of Q20, LEADING removed bases at the front of the read below the set threshold of Q10. The quality threshold at the front of the read was set to a less stringent threshold because reads trimmed at the front can no longer be de-multiplexed based on the barcode (i.e. different barcoded individuals could no longer be distinguished) and thus would have to be discarded. Q10 has a probability of 1 error in 10 bases, which would translate into one error in the barcode. All barcodes differ by 2 nucleotides and thus a barcode containing one error can still be safely attributed to a specific sample. MINLENGTH was set to 36 bases as reads below this length would no longer be informative.

Cleaned data were then imported into CLC Genomics workbench version 7.0 (CLC Inc., Aarhus, Denmark) and de-multiplexed by barcode. For library set 1, reads were de-multiplexed using P1-indices only. P2 indices consisted of a mixture of two different adapters in roughly equal ratios and were trimmed off sequencing reads. For eight samples in set 2, the presence of both P1 and P2 indices was crucial to positively assign barcodes back to the sample. All ungrouped sequences (sequence pairs which could not be assigned back to samples based on P1 and P2 barcodes) were de-multiplexed again based on the P1 indices only for those 16 samples that had a unique P1 barcode.

Velvet Assembly

Popular pipelines such as Stacks (Catchen et al. 2011) or PyRAD (Eaton 2013) which are often used to assemble or cluster data are at present unsuitable to deal with samples that are polyploid. Thus, to prepare our data for downstream ploidy analysis, we decided to assemble our raw data outside of these pipelines. Paired-end reads were assembled separately for each species using Velvet version 1.2.10 (Zerbino & Birney 2008). The wrapper script Velvetoptimiser version 2.2.5 (Zerbino 2010; Gladman & Seeman 2012) was used to optimize the three parameters: k (word length), expected coverage and coverage cutoff. The optimisation function used was the default N50. In cases where assembly was difficult, the

optimisation function was changed to number of contigs, optimizing for number of contigs instead of for length.

Velvet uses coverage to distinguish unique regions of the genome from repetitive regions (Zerbino 2010; Zerbino & Birney 2008). Thus, high variance in coverage can cause errors in the assembly. Because RAD data coverage is expected to be highly non-uniform and k-mer coverage in Velvet was highly variable after first optimization attempts, we normalized the coverage across contigs using the digital normalization script bbNorm which is part of the bbMap package (http://sourceforge.net/projects/bbmap/).

For each species, reads of all samples (including replicates) were combined for the assembly. As reverse reads vary more in coverage (due to the random shearing step) as well as to further facilitate the assembly of repetitive regions, reverse reads were assembled in a separate run using a longer kmer length. These contigs were then passed back into the complete assembly as long reference reads. We used a smaller k for this step to aid connecting the forward and reverse reads.

Mapping

To estimate the quality of the assembly, raw reads for all species were mapped back to the contigs created by Velvet using the BWA-mem algorithm (Burrow-Wheeler-Alignment) (Li 2013). A contig was considered useable if both forward and reverse read of a read-pair map back to the same contig. These 'verified' contigs were then used for further downstream analyses. The mean read depth was calculated using the GATK Depth of Coverage Tool (Mckenna et al. 2010).

PyRAD Analysis

As polyploidy can occur in connection with hybridization events (allopolyploidy), we constructed a phylogeny in order to detect potential conflicts within the RAD data set, as well as conflicts between a RAD based phylogeny and the mitochondrial based phylogeny from Alexandrou (2011)**.** As the velvet assembly may be prone to assemble paralogs, we constructed a conservative data set using the pipeline PyRAD (Eaton 2013)**.** PyRAD filters out potential paralogous sequences by identifying sequences with more than a set number of heterozygous sites (default of 5) and with a heterozygous site shared between a set number of samples (default of 3). PyRAD also discards clusters with more than 2 haplotypes in default

mode as it is designed for diploid organisms. We ran PyRAD in default mode, with data within species clustered at 90% similarity, and these were then clustered at 80% across species.

Appropriate models of nucleotide substitution for the RAD data set were determined using jModeltest (Guindon & Gascuel 2003; Darriba et al. 2012). Maximum likelihood and Bayesian phylogenetic analysis were conducted in MrBayes (Ronquist et al. 2012) and RAxML (Stamatakis 2014) using the concatenated PyRAD output using the nucleotide substitution model suggested by jModeltest. Two separate MCMC runs were conducted in MrBayes and run for 1,000,000 generations with random starting trees. We executed 1000 Rapid Bootstrap searches in RAxML using the Rapid Bootstrapping Algorithm.

To estimate changes in ploidy level within the Corydoradinae we used two different complementary methods: (i) haplotype frequencies in each contig and (ii) read count ratios for bi-allelic SNPs.

Ploidy Analysis I – Haplotype Number per contig

In this analysis, we quantified the number of different haplotypes for each contig. The filtered Bam files were run through Hapler which performs haplotype calling in low-diversity, low-coverage short-read sequence data (O'Neil & Emrich 2011; O'Neil & Emrich 2012). In order to maximise coverage (which is crucial to identify haplotypes and distinguish them from sequencing errors), individual replicates were merged prior to the analysis. As haplotype assembly can be complicated by reads mapping to consecutive stretches of DNA that do not fully overlap, the data were also filtered to include only haplotypes with a minimum of 20 reads and exclude all alignments that stretch beyond 200 bases. Hapler can base haplotypes on SNPs called internally or using a provided SNP list. As the SNPs used for the Ratio-Analysis were filtered very conservatively, we utilized Hapler's internal SNP-caller in the binomial mode. Hapler output was processed using custom scripts and MS Excel. Data were visually inspected using Tablet (Milne et al. 2013). Contigs were grouped according to haplotype number and frequencies were calculated.

We used general linear models to identify significant changes in haplotype number frequencies across species. We first fitted a fully saturated model, and then removed haplotype category from a second updated model. A subsequent Analysis of Deviance was

conducted on both models, to detect significant effects of either species or haplotype category. All analyses were conducted in Rstudio (RStudio 2012).

Ploidy Analysis II –SNP Frequency

To estimate changes in ploidy level within the Corydoradinae, we calculated the read count ratios for bi-allelic SNPs as outlined in Yoshida et al. 2013. Briefly, this is based on the principle that, mean read ratios for each bi-allelic SNP are expected to be roughly ½ and ½ in a diploid. In a triploid, read ratios would be expected to be $^{1}/_{3}$ and $^{2}/_{3}$ and in a tetraploid either ¼ and ¾ or ½ and ½ respectively.

Here, only putatively coding regions were used to avoid noise from multi-copy repetitive regions where SNP ratios are not predictable. Thus, Repeatmasker version 4.0 (Smit et al. n.d.) was used to mask assembled contigs for each species. Blastx (Camacho et al. 2009) was then used to identify potential coding regions. Raw reads were mapped to these identified candidate-coding regions using BWA-mem with the same settings as outlined above. Freebayes was then used (Garrison & Marth 2012) to call polymorphisms with a minimum occurrence of ten reads. Reads from individual replicates are typically combined at this stage to increase coverage. Here however, Freebayes was run on each replicate individually to allow SNP verification. For each species, resulting datasets were further quality filtered to contain only bi-allelic SNPs, a maximum total depth of 300 and a minimum reference-allele count of 5 per individual replicate. SNPs shared between both replicates were considered real and read counts for the reference and alternate SNPs of both replicates were combined. Histograms of SNP read ratios (reference reads divided by total reads, and alternate reads divided by reference reads) for each individual were created using ggplot2 in R studio (Wickham 2009; RStudio 2012).

We were unable to model expected frequencies for autopolyploid individuals due to inadequate model assumptions: our read-frequency data shows a clear peak around 0.5 for most species, indicating preferential pairing of chromosomes during meiosis. Modelling an autopolyploid according to coalescent theory however would require that homeologous chromosomes are completely exchangeable and that there is no preferential pairing (Arnold et al. 2012; Arnold et al. 2015). Furthermore, it is also possible that these species are allopolyploid, or segmental polyploids, which would lead to a mixture of bivalent and

multivalent formation during meiosis and thus an unpredictable mixture of different ratios at different contigs.

Instead, we compared frequencies of different species against each other to detect significant changes within the phylogeny. We explored these differences statistically by dividing the data into three bins (0.25, 0.5 and 0.75) and performed pairwise Chi-Square tests in R. P values were Bonferroni-corrected to account for multiple comparisons.

3.3 Results

Sequencing and Data Cleaning

The first sequencing run yielded *c.* 104 million paired reads (GC content 47%). After quality filtering and trimming, 93.52% of the original sequences remained. The second sequencing run resulted in *c.* 117 million paired sequences (GC content 46%), but was of lower quality with only 81.99% of paired sequences surviving. The overall quality of sequencing run 2 was very high except for a dip in quality after 70 bases in the reverse reads. Cleaned and trimmed reads were de-multiplexed, with number of reads per barcode ranging from 1.4 million (*A. poecilius*1-Replicate1) to 20 million (*C. metae*2-Replicate1) as visualized in figures 8 and 9.

Figure 8. The number of reads in millions obtained for each sample in sequencing run 1.

Figure 9. The number of reads in millions obtained for each sample in sequencing run 2.

Velvet Assembly

Results for the assembly are listed in table 7. The number of contigs assembled for each species ranged between 13166 (*C. aeneus*) and 58604 (*C. imitator*). The assembly yielded fewer contigs for species in the second sequencing run, due to shorter reverse reads. This is reflected in a lower N50 distribution (270-355) when compared to the first sequencing run (397-447). The number of contigs assembled was also higher for species in sequencing run 1 (28k-58k in run 1 compared to 13k-32k in run 2). We only used contigs for downstream analyses if both forward and reverse reads of a read-pair map to the same contig. These are listed as verified contigs in table 7.

Differences between the sequencing runs can also be seen here (table 7). In general, the discrepancy between contigs assembled and contigs to which paired reads map was small, indicating that the velvet assembly worked well. However, for sequencing run 2, a much lower percentage of sequences mapped back in verified pairs, while percentage of single reads mapping back to the same contigs was comparable to sequencing run 1. This indicates that velvet failed to connect reverse and forward reads in many cases for sequencing run 2, again likely due to the shorter reverse reads. While the quality drop in the reverse reads of the second sequencing run resulted in a less successful assembly for the sequenced species

(lineages 2 to 7), analyses based on successfully assembled contigs (see haplotype analysis and SNP analysis) should otherwise not be affected.

Table 7. Number of reads for all species and replicates, as well as basic statistics from the Velvet assembly. The number of verified contigs represents the number of contigs that were used for downstream analyses.

Species	**Velvet-Contigs**	**N50**	**Mean Depth**	**Remaining Contigs after Filter**	**C- value**
Megalechis sp.	37345	432	45.26	**36426**	1.58
Corydoras fowleri	28047	414	40.39	**26743**	0.65
Aspidoras poecilius	20081	299	97.675	**18374**	0.76
Scleromystax kronei	24100	335	41.025	**20866**	0.8
Corydoras pygmaeus	34490	355	116.18	**30051**	2.68
Corydoras elegans	22238	290	49.63	**18674**	2.24
Corydoras nattereri	13166	270	49.625	**11972**	1.79
Corydoras aeneus	19375	292	55.875	**17787**	1.84
Corydoras imitator	58604	447	54.825	**49552**	2.3
Corydoras metae	44052	397	104.175	**40763**	4.15
Corydoras araguaiaensis	42224	403	145.865	**40026**	4.36

PyRAD Analysis

PyRAD identified between ~33k contigs (*C. fowleri 2*) and ~90k contigs (*C. araguaiaensis*) across species with a minimum coverage of ten (table 8). Mean coverage varied between 32 in *C. nattereri 1* and 176 in *C. pygmaeus 1*. PyRAD also filters out RAD-tags that appear to be potential paralogs, based on estimated heterozygosity and error rates. In the outgroup and the basal lineage 1 *C. fowleri* samples (C-value = 0.65pg), roughly about 4% of RAD tags are discarded due to paralog-filters. The loss doubled from lineage 2 onwards and increased to 51% in the lineage 9 individual *C. araguaiaensis 1 (*C-value = 4.32pg*).*

The consensus data set created across all species consisted of 3004 loci, with 1015 loci present in *Megalechis sp.* and all other species containing at least 1800 loci of the common data set. The phylogenetic analysis based on the concatenated dataset recovered only one topology both for the Bayesian as well as the Maximum likelihood analysis (see figure 10).

As the PyRAD analysis was based on forward reads only, no difference between sequencing runs became apparent.

Table 8. Basic statistics from the PyRAD analysis for all species

Lineage	Sample	C- value	RAD-Sites after coverage filter	Mean Coverage after filter	RAD-Sites after paralog filter	% filtered	% Polymorphic
0	*Megalechis*	1.58	37,628	64.67	36,256	3.65	0.16
1	*C. fowleri 2*	0.65	32,765	45.63	31,436	4.06	0.13
1	*C. fowleri 1*		33,040	48.34	31,515	4.62	0.19
2	*A. poecilius 1*	0.76	35,930	74.03	32,576	9.33	0.28
2	*A. poecilius 2*		42,925	113.79	39,046	9.04	0.28
3	*S. kronei 1*	0.8	37,199	70.25	31,659	14.89	0.41
3	*S. kronei 2*		36,767	65.81	33,512	8.85	0.35
4	*C. pygmaeus 1*	2.68	60,292	176.06	54,816	9.08	0.29
4	*C. pygmaeus 2*		55,802	154.91	51,930	6.94	0.30
5	*C. elegans 1*	2.24	74,679	54.63	63,323	15.21	0.46
5	*C. elegans 2*		77,720	56.73	66,300	14.69	0.47
6	*C. nattereri 1*	1.79	66,457	32.20	54,451	18.07	0.48
6	*C. nattereri 2*		69,654	35.92	56,735	18.55	0.49
7	*C. aeneus 1*	1.84	48,019	57.48	39,584	17.57	0.38
7	*C. aeneus 2*		50,877	59.42	43,037	15.41	0.35
8	*C. imitator 1*	2.3	61,084	61.75	55,020	9.93	0.24
8	*C. imitator2*		61,962	69.23	55,583	10.30	0.23
9	*C. metae 1*	4.15	63,540	53.79	53,237	16.21	0.57
9	*C. metae 2*		76,613	63.54	60,192	21.43	0.42
9	*C. araguaiaensis 1*	4.36	87,170	60.24	42,170	51.62	0.55
9	*C. araguaiaensis 2*		89,921	59.25	61,819	31.25	0.42

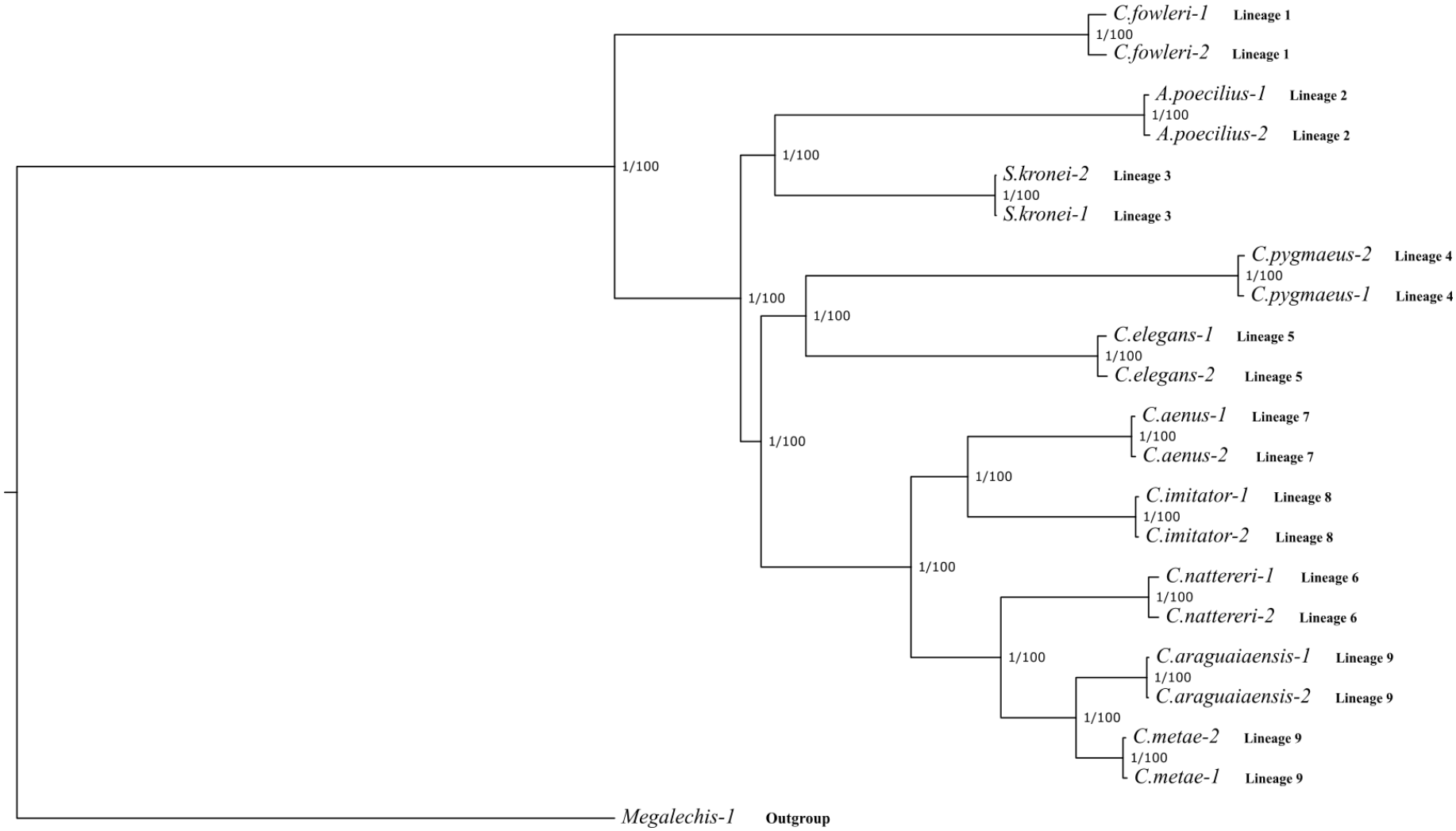


Figure 10. Topology recovered from Maximum Likelihood analysis using RAxML and Bayesian phylogenetic analysis using Mr bayes. Node values represent posterior probabilities/bootstrap support.

Ploidy-Analysis 1 – Number of Haplotypes per Contig

In an ideal assembly, we would expect each contig to be representative of one homologous genomic location and thus possess a maximum of two haplotypes in a diploid. However, during the assembly of raw reads into contigs, assemblers often fail to distinguish between regions of high similarity (for instance recently duplicated regions), which may be mis-assembled into one contig. In a paleo- or neopolyploid, we would expect a higher frequency of these paralogous contigs, i.e. contigs that are representative of more than one homologous genomic region. Thus, the number of haplotypes mapping to each contig created in the Velvet assembly were quantified. With every additional WGD event that may have occurred, the number of highly similar, duplicated paralogous regions in the genome doubles. Depending on the age of the event and the degree of re-diploidization and divergence of duplicates, we would expect an increase of paralogous regions to manifest itself in a detectable increase in contigs with multiple haplotypes.

The proportion of contigs with one haplotype, two haplotypes or multiple haplotypes is shown in figure 11. There are large differences in the number of haplotypes per contig across species. For *Megalechis* (outgroup) and *C. fowleri* (lineage 1) ~ 75-80% of contigs have only one haplotype, this drops sharply to ~45% from *A. poecilius* (lineage 2) to *C. aeneus* (lineage 7). Similarly, the number of contigs with two haplotypes doubles from ~ 16% in *C. fowleri* (lineage 1) to 31% in *A. poecilius* (lineage 2), with 3x the proportion of contigs with multiple haplotypes. *C. imitator* (lineage 8) has a higher proportion of contigs with one haplotype in comparison with lineages 2-7, although the number of multi-haplotype contigs remains roughly similar. Finally, *C. metae* and *C. araguaiaensis* (both lineage 9) have only ~25% of contigs with a single haplotype and with the majority of contigs having two or multiple haplotypes. Such patterns indicate one or several WGD events following the splitting of lineages 1 and 2. Remarkably, genome sizes of *A. poecilius* and *S. kronei* are highly similar to that of *C. fowleri* which had previously led to the assumption that WGD events would have occurred after the lineage 3-lineage 4 split. This could be due to a higher degree of re-diploidization in comparison with earlier lineages, or perhaps could be indicative of a hybridization/allopolyploidy event, as we may expect sufficient divergence in parental genomes to avoid misassembly.

The Analysis of Deviance on the fully fitted and updated glm models on the haplotype count data supports a significant change ($p<0.001$) in the number of contigs belonging to different haplotype categories across species, with a significant interaction between haplotype

number and species. Thus, proportions vary significantly both with species as well as with haplotype category.

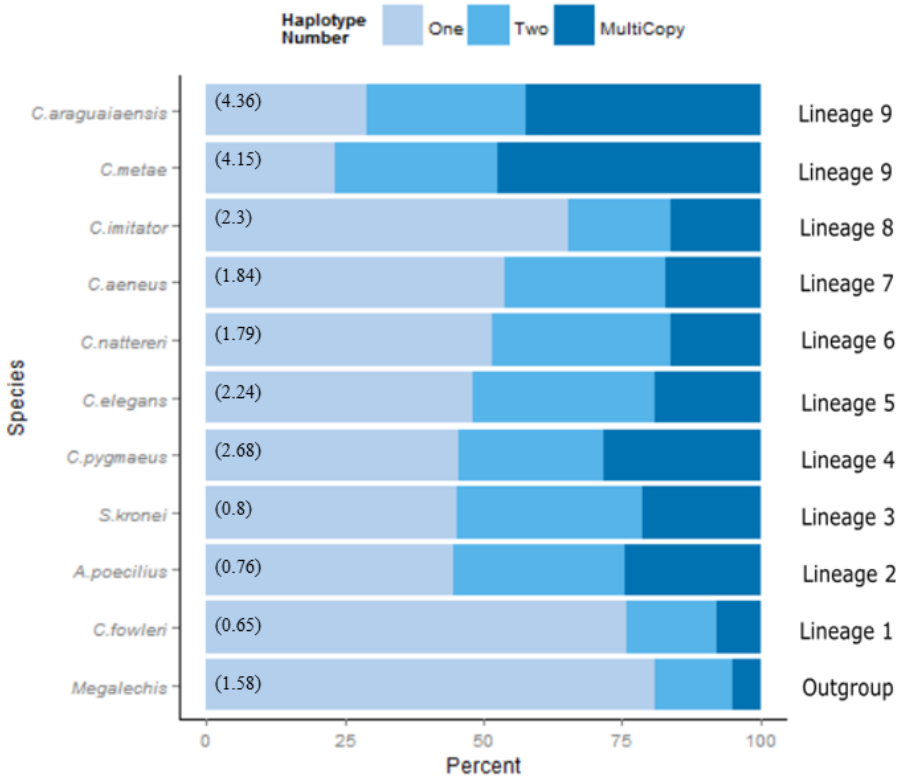


Figure 11. Proportion of assembled contigs with one, two or multiple haplotypes for each species. Genome size for each species is shown in brackets (C-value).

Ploidy-Analysis 2 – SNP frequency histogram shape

In SNP frequency analysis, both *Megalechis* (outgroup) and *C. fowleri* (lineage 1) display a clear peak around 0.5, i.e. the majority of bi-allelic SNPs have roughly an even read number as expected in a diploid species (figure 12). Species in lineages 2-8 all display a large peak at 0.5 with slight differences in shape that could reflect differences in paralog retention after genome duplication events. Most species also display slightly raised peaks at 0.25 and 0.75 frequencies which could be a sequencing artefacts or alternatively reflect paralogs, which would be found even in a diploid genome.

Differences between the shapes of the distributions in are not statistically different among lineages 1-3 and *Megalechis* (table 9), suggesting that all four species are functionally diploid. All samples representing lineages 4-8 differ significantly from all others, with the exception of *C. aeneus* (lineage 7) and *C. nattereri* (lineage 6).

C. metae shows a markedly broader distribution, with increased 0.25 and 0.75 peaks that are only slightly shorter than the 0.5 peak. In *C. araguaiaensis*, the 0.25, 0.5 and 0.75 peak are roughly equal and form a broad plateau. In a perfect tetraploid, SNPs should display either at a 0.5 read ratio or a 0.75/0.25 ratio. Thus, *C. araguaiaensis* and *C. metae* display SNP frequency distributions which are consistent with being functionally tetraploid. *C. metae* does not differ significantly from either *C. aeneus* or *C. nattereri* (after Bonferroni correction for multiple testing), with both of these species also showing raised 0.25 and 0.75 peaks and consistent with being tetraploids that are in the process of re-diploidizing. Similarly, the gradient in SNP read ratio distributions between lineages 1-8 could be a sign of different stages of re-diploidization after potential WGD events that are indicated in the haplotype analysis.

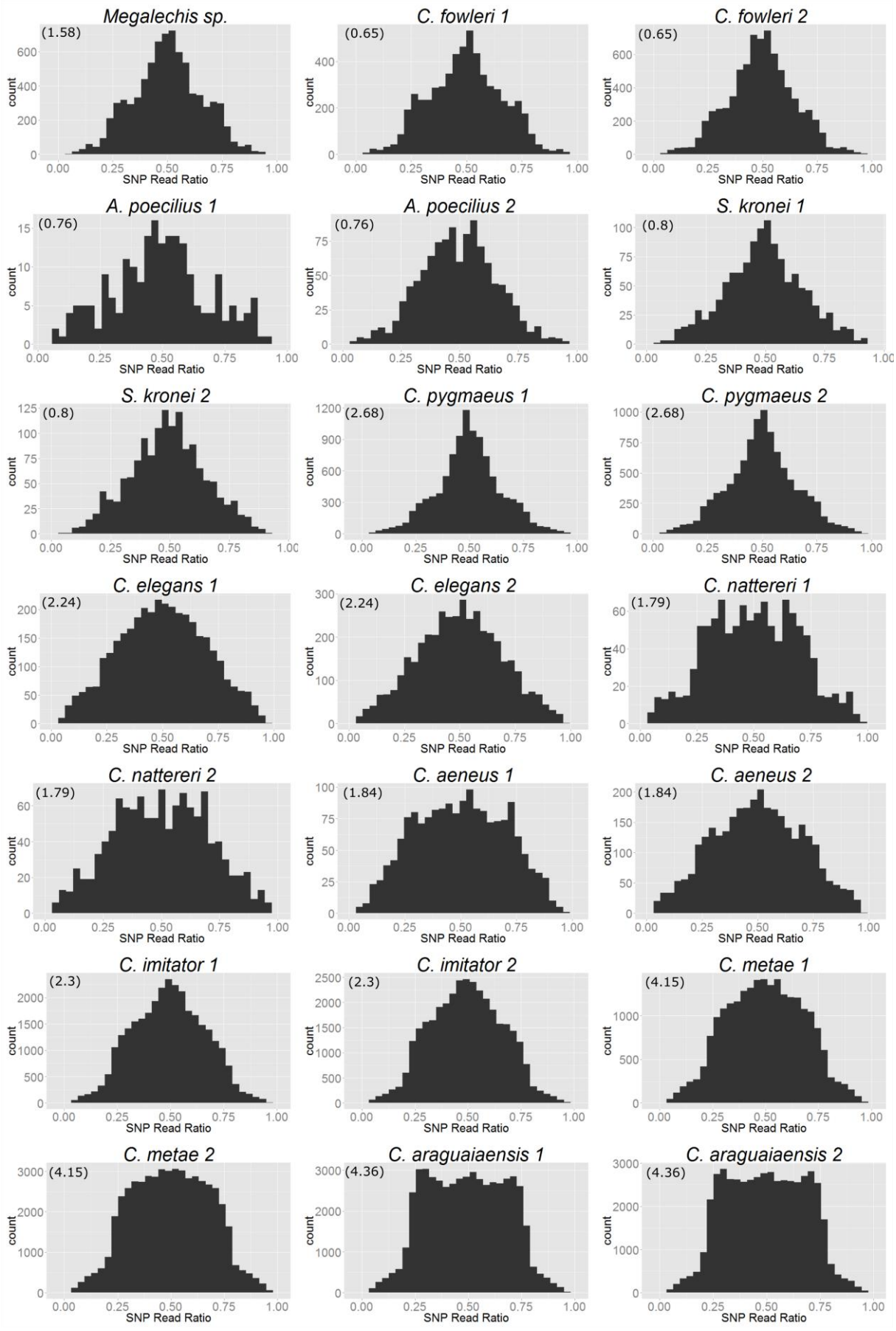


Figure 12. Frequencies of bi-allelic SNP read ratios for all 21 samples. Genome size (C-value) displayed in brackets.

Table 9. P-values for all pairwise chi-square comparisons of 0.25, 0.5 and 0.75 SNP read ratio frequency bins. The ***** symbol indicates that significance does not hold after Bonferroni correction.

	Megalechis	*C. fowleri*	*A. poecilius*	*S. kronei*	*C. pygmaeus*	*C. elegans*	*C. nattereri*	*C. aeneus*	*C. imitator*	*C. metae*
Megalechis										
C. fowleri	0.242									
A. poecilius	0.328	0.5395								
S. kronei	0.2504	0.4264	0.9742							
C. pygmaeus	<0.001	<0.001	<0.001	<0.001						
C. elegans	<0.001	<0.001	<0.001	<0.001	<0.001					
C. nattereri	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001				
C. aeneus	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.3755			
C. imitator	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001		
C. metae	<0.001	<0.001	<0.001	<0.001	<0.001	0.001777*****	0.01437*	0.1927	<0.001	
C. araguaiaensis	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001

3.4 Discussion

In this study, we successfully identify significant shifts in multi-haplotype contigs as well as in bi-allelic SNP read ratios, which indicate that at least two major duplication events occurred in the Corydoradinae. Previous studies investigating genome size and chromosome number changes across the group suggest that lineages 1-3 are diploid, with genome size increasing from lineage 4-9. Here, we find that contrary to previous predictions based on genome size, a WGD event likely occurred prior to the split of lineages 2 and 3, and that a further WGD event occurred in lineage 9. We employed two different ways to quantify paralogs across species. First, we identified bi-allelic SNP read count ratios, a technique which has recently been used to determine ploidy levels in Yoshida et al.(2013) and in Arnold et al. (2015). Second, we present a novel approach for detecting potential WGD using a RAD data set by directly quantifying the number of haplotypes mapping to our assembled contigs.

In addition to quantifying paralogs, we successfully resolved the phylogenetic relationships of the Corydoradinae using nuclear markers and thus demonstrate that RAD data can be used both to resolve phylogenies of highly diverged lineages as well as to make inferences about the evolutionary dynamics of genome evolution of non-model species.

Evidence for WGD events across the Phylogeny

One of the biggest challenges in working with polyploid species is the inevitable mis-assembly of homeologous and repetitive parts of the genome into one paralogous sequence. While sequencing read lengths are steadily increasing, reads are still often too short to distinguish between homeologous/paralogous regions which leads to errors in assembly and makes accurate downstream genotyping difficult. Paralogs in diploids are often removed during the analysis (Mastretta-Yanes et al. 2014; Hohenlohe et al. 2012), even though they can be highly informative both for population genomics as well as in an evolutionary context (Mastretta-Yanes et al. 2014). The assembly may be complicated further by the often varying degree of re-diploidization and divergence of paralogs after WGD events, leading to copy number variation for different parts of the genome (Wolfe 2001). In the present study, we quantified paralogs instead of attempting to resolve or remove them.

After our paired-read assembly, we expected a large number of assembled paralogous contigs with multiple haplotypes. With 75% and 80% of contigs in *Megalechis* and *C. fowleri* respectively only displaying one haplotype, it appears that velvet correctly assembled contigs for the majority of our relatively short RAD-data. In fact, determining the number of haplotypes per assembled scaffold/contig could serve as a useful quality control of assembled data in the absence of reference genomes or ploidy-information. Based on the haplotype analysis and the read count ratio SNP-analysis, *C. fowleri* and *Megalechis* appear functionally diploid: For the majority of contigs only one or two haplotypes per individual were identified, and the SNP read ratios show a clear peak around 0.5. This is also the case for *A. poecilius* and *S. kronei*. Surprisingly, there is a large shift in haplotype number proportions between *C. fowleri* and *A. poecilius*/*S. kronei* that indicate an ancient duplication event, despite the genome sizes of lineages 1-3 being nearly identical. Based on genome size data (Alexandrou 2011), we assume that *A. poecilius* and *S. kronei* have re-diploidized but still retain an increased number of multi-copy regions. Initially, in a polyploidy event, similarity between duplicated regions is maintained through bivalent formation during meiosis, allowing for recombination between all copies. The key step in the re-diploidization of a polyploid individual is the return to bivalent formation of chromosomes during meiosis over time, with duplicated copies located on different chromosome pairs gaining in divergence (Wolfe 2001). If the majority of the genome has returned to a re-diploidized state and as such become a paleopolyploid, we would expect the majority of contigs to have one or two haplotypes and the majority of bi-allelic SNPs to occur at a 0.5 ratio. While the most prevalent fate of duplicated regions is subsequent loss, a significant number of duplicate genes appear to be retained in fish and plant paleopolyploids. There is no additional shift in haplotype numbers for species representative of lineages 4 to 7, though all appear significantly different to each other in terms of SNP read ratios and show an increase in genome size. These differences between species could represent different stages of re-diploidization. As RAD-Seq data has been shown to underestimate true haplotype diversity (Arnold et al. 2013), it is also plausible that we may not have enough data to pick up additional duplication events.

The haplotype analysis as well as the SNP read count ratio analysis detected an observable shift between lineages 1 and 2-8 as well as in lineage 9. This shift is similarly apparent in the PyRAD data when comparing the amount of RAD-tags lost to paralogs and the number of polymorphisms identified. Species in lineage 2 and 3 thus likely represent paleopolyploids. With half of all SNPs occurring at a 0.25/ 0.75 ratio, it seems plausible that *C. metae* and *C. araguaiaensis* could be functionally tetraploid, i.e. observed ratios could be

explained by tetravalent formation and recombination during meiosis. The haplotype analysis for *C. imitator* shows that the majority of contigs contain only one or two haplotypes, an increase in comparison to lineages 2-7 and more comparable to lineage 1. This could be explained by a higher degree of re-diploidization. Alternatively, *C. imitator* could be a segmental polyploid or allopolyploid. Allopolyploids form as a result of the hybridization of different species, but with two more divergent genomes compared to an autopolyploid, resulting in chromosomes often forming bivalents and not multivalents. Depending on the degree of divergence of the parental species, there is a mixture of bivalent/multivalent formation of chromosomes. Lineages 4-7 likely represent different stages of re-diploidization and may have undergone additional WGD events for which our data was not powerful enough to detect them. *C. aeneus* and *C. nattereri* in particular appear similar to *C. metae* in the SNP read ratio analysis, though they fail to show a significant shift in terms of the haplotype analysis.

Generally, our analyses emphasize the value of determining haplotype number per contig as we have done, as the bi-allelic SNP read count ratios did not pick up the clear shift between lineages 1-3, most likely due to re-diploidization.

Due to sequence length limitations of the RAD-data, we cannot show that paralogs arose at the same time and are thus the result of WGD, as done elsewhere using synonymous substitution rates (Blanc & Wolfe 2004) . An alternative method would be a phylogenetic or topology based approach (Wolfe 2001; Jiao et al. 2011), though there is a very small overlap in orthologous gene families across species due to their high divergence. Thus, we cannot exclude the possibility that the increase in paralogs is the result of other duplication mechanisms, such tandem, segmental or chromosome duplications, which can be difficult to distinguish from Whole Genome Duplication events particularly when faced with old events and large scale re-arrangement, loss and mutation (Durand & Hoberman 2006) . However, the RAD data indicates that there is a 30% shift between lineages 1 and 2 in terms of haplotype proportions. Based on previous phylogenetic analysis, lineages 1 and 2 split at least 30 million years ago (Alexandrou 2011), and a 30% retention of duplicates would be in line with other paleopolyploid species (Blanc & Wolfe 2004; Ludwig et al. 2001) and presents the most parsimonious explanation.

Phylogenetic Analysis

The comprehensive phylogenetic framework presented in Alexandrou *et al.* (2011) was based on mtDNA with a single nuclear gene included. In this study we used the RAD data to construct a nuclear phylogeny to identify whether there were significant differences in the order of lineages in the nuclear vs. mtDNA trees. As the haplotype data indicated that a large number of contigs appear to conform to diploid status, we ran our forward reads through the PyRAD-pipeline for phylogenetic purposes, as has been done previously for other polyploid species (Ogden et al. 2013).

The phylogeny was concordant with the mtDNA tree with one exception. *Corydoras nattereri* (lineage 5) formed a monophyletic clade with the two lineage 9 species, *C. araguaiaensis* and *C. metae)*. Based on the mitochondrial phylogeny, we would expect *C. nattereri* to be basal to *C. aeneus* and *C. imitator (*lineages 7 and 8). The reasons for this conflict could be error in the phylogenetic analysis as a result of insufficient resolution or information in the mitochondrial-based analysis, incomplete lineage sorting or introgression and hybridization events. The SNP read ratio analysis revealed that *C. nattereri* could be a polyploid, which together with the mitochondrial conflict, could be suggestive of a hybridization event. This conflict between nuclear and mitochondrial phylogeny may only affect *C. nattereri* and not be representative of the entire lineage 6. More data for additional species would be necessary to identify the cause of the conflict.

Conclusion

It is clear that the evolutionary history of the Corydoradinae has been significantly influenced by large-scale duplication events. The increase in genome size across the Corydoradinae led to the hypothesis that WGDs could have occurred within this species-rich subfamily. While we believe that we have found conclusive evidence for several WGDs, these cannot explain the variation in genome size on their own. Further studies are needed to identify potential lineage-specific events as well as other mechanisms that have potentially driven genome size in the Corydoradinae. Such alternative mechanisms could include a Transposable Element expansion (see chapter 5), pseudogene and intron accumulation as well as chromosome-level events (Gregory 2005a). With several rounds of WGD shared by a large number of diverse species, the Corydoradinae also make an excellent model system to study the retention of gene duplicates and the mechanisms of re-diploidization in unprecedented

detail at different evolutionary timescales: The FSGD occurred ca 350 million years ago (Glasauer & Neuhauss 2014) and thus comparing gene retention in extant teleosts today only gives us a snapshot of the mechanisms impacting the genome after a WGD duplication event. The Corydoradinae provide a large species sample size and multiple WGDs at different time scales at different stages of re-diploidization. This makes the Corydoradinae a unique and exciting model system for studying the genome dynamics within vertebrates in general.

3.5 Supporting Information

Table 10. RAD reads per sequenced sample and putative contigs assembled in Velvet. Sequence reads were then mapped back to putative contigs. Only those contigs with properly paired mapped reads were kept for downstream analyses.

Lineage	Species	Species	#Reads	Putative Contigs	% Reads mapping back to own contigs Properly paired (total reads mapped)
0	*Megalechis*	Mega1-R1	3354110	37345	84.95 (93.12)
0		Mega1-R2	2996048		81.55 (89.88)
1	*C. fowleri*	Fow1-R1	2360806	28047	69.88 (85.97)
1		Fow1-R2	2691040		70.47 (85.72)
1		Fow2-R1	2707300		70.26 (85.1)
1		Fow2-R2	2536026		69.5 (84.89)
2	*A. poecilius*	Apo1-R1	1381450	20081	67.54 (76.93)
2		Apo1-R2	5239442		68.46 (78.13)
2		Apo2-R1	5148758		68.22 (77.73)
2		Apo2-R2	7245580		67.65 (77.32)
3	*S. kronei*	Skro1-R1	2252998	24100	55.53 (75.48)
3		Skro1-R2	4027790		57.56 (76.94)
3		Skro2-R1	3061372		55.79 (74.92)
3		Skro2-R2	3060050		58.03 (77.15)
4	*C. pygmaeus*	Pyg1-R1	11511440	34490	64.19 (86.86)
4		Pyg1-R2	11281860		64 (86.76)
4		Pyg2-R1	12146160		61.16 (85.26)
4		Pyg2-R2	11147618		60.57 (84.47)
5	*C. elegans*	Eleg1-R1	8614162	22238	26.59(67.92)
5		Eleg1-R2	6496120		25.48 (65.71)
5		Eleg2-R1	8770490		27.18 (67.01)
5		Eleg2-R2	5768858		28.54 (66.85)
6	*C. nattereri*	Nat1-R1	3750502	13166	24.26(66.80)
6		Nat1-R2	5374010		23.59 (66.65)
6		Nat2-R1	6109870		23.62 (65.52)
6		Nat2-R2	4566728		23.84 (66.44)
7	*C. aeneus*	Aen1-R1	7987504	19375	29.98 (78.83)
7		Aen1-R2	2845866		24.10 (77.77)
7		Aen2-R1	5253556		28.83 (78.48)
7		Aen2-R2	6622462		31.03 (78.12)
8	*C. imitator*	Imi1-R1	7583986	58604	61.99 (95.19)
8		Imi1-R2	8033302		62.52 (94.59)
8		Imi2-R1	8718744		62.52 (94.78)
8		Imi2-R2	9132258		61.61 (95.10)
9	*C. metae*	Me1-R1	11187184	44052	45.72 (95.67)
9		Me1-R2	13057302		45.25 (95.03)

9		Me2-R1	20102230		46.35 (94.64)
9		Me2-R2	16330000		48.63 (94.65)
9	*C. araguaiaensis*	Ara1-R1	15460804	42224	66.75 (93.25)
9		Ara1-R2	13187802		66.10(92.73)
9		Ara2-R1	13055310		65.86(93.54)
9		Ara2-R2	15617158		66.68(92.92)

Chapter 4 - Paralog Retention following large scale duplication events in the Corydoradinae

"Because of natural selection, organisms have been able to adapt to changing environments, and by adaptive radiation many new species were created from a common ancestral form. Yet, being an effective policeman, natural selection is extremely conservative by nature. Had evolution been entirely dependent upon natural selection, from a bacterium only numerous forms of bacteria would have emerged. The creation of metazoans, vertebrates and finally mammals from unicellular organisms would have been quite impossible with previously non-existent functions. Only the cistron which became redundant was able to escape from the relentless pressure of natural selection, and by escaping, it accumulated formerly forbidden mutations to emerge as a new gene locus."

Preface, Susumu Ohno, Evolution by Gene duplication, May 1970

4.1 Introduction

The duplication of genes has long been suspected to play a fundamental role in the creation of novel genes and gene functions (Ohno et al. 1970; Kondrashov et al. 2002; Taylor & Raes 2004; Rensing et al. 2007). A number of different mutative mechanisms can lead to the duplication of genes. These include tandem duplication events, the duplication of individual chromosomes (aneuploidy), Whole Genome Duplication (WGD) events, as well as duplicative transposition (the duplication and movement of genomic fragments through Transposable Elements) (Taylor & Raes 2005). Gene duplications are surprisingly common, occurring at rates comparable to those of nucleotide substitution (Lynch & Conery 2000; Lynch & Conery 2003). Despite this surprisingly high rate of gene duplication, most duplicated genes are lost from the genome within a few million years as a result of silencing and the accumulation of deleterious mutations. Genes that are retained may undergo sub-functionalization (original function is divided between the duplicate copies) or neo-functionalization (the acquisition of a new function) (Lynch & Conery 2000) or continue to be expressed in duplicate.

There is increasing evidence that the evolutionary fate of duplicate gene copies is non-random and is heavily biased by gene function and duplication mode. Lynch and Conery (2000) noted that even though the overwhelming majority of duplicate genes should be lost relatively quickly, gene duplicates resulting from a WGD events (also termed ohnologs) have been retained in much larger numbers than expected. For instance, it is estimated that approximately 15% of gene duplicates have been retained in teleosts since the teleost specific duplication event (Braasch & Postlethwait 2012), and roughly 20% being retained in the two paleopolyploid yeast species *Sacchromyces cerevisae and Sacchromyces castellii* (Byrne & Wolfe 2007). Moreover, there appears to be a distinct difference in the type of genes that survive in duplicate after WGD events when compared with small-scale duplications or transposition events. This has been attributed in part to gene dosage balance: after a genome duplication event, genes that co-operate and are co-regulated are retained by purifying selection to maintain dosage balance (Freeling 2009; Conant et al. 2014; Gout & Lynch 2015). Other genes appear to be "duplication resistant" across a diverse range of taxa and have quickly reverted back to single copy status after independent WGD events (Paterson et al. 2006).

While many models have been proposed to explain gene retention and subsequent evolution after duplication **(**reviewed in Innan & Kondrashov 2010; Conant et al. 2014) a common feature is a focus on the increased freedom of these duplicates to evolve under reduced purifying selection which in some cases may lead to adaptive evolution. Several studies have shown that paralogs in teleosts evolve more quickly than their single-copy orthologs in mammals (Wagner et al. 2005; Brunet et al. 2006; Crow et al. 2009). Gene duplicates in rodents appear to have accelerated positive selection in the first 12 million years, before duplicates return to rates similar to pre-duplication level (Pegueroles et al. 2013). Many important gene families such as Haemoglobin and Immunoglobin are the product of gene duplications (Taylor & Raes 2005) and some gene families such as the *Hox* (chapter 2) have played important roles in elucidating the role of ancient WGDs in the evolution of the vertebrates (Prohaska & Stadler 2004; Amores et al. 1998; Crow et al. 2006).

While it has clearly been demonstrated that duplicated genes can have an evolutionary impact and lead to the creation of novel features, the link between WGD or polyploidization, the evolution of new traits and the diversification of species is less clear (Otto 2007)**.** In part, this is due to a paucity of quantitative studies that directly compare closely related groups that have undergone the same WGD event. Two studies that have addressed such comparisons: Morel et al. (2015) demonstrated that the reciprocal loss of duplicate genes may be a powerful tool in facilitating adaptation and increasing biodiversity in yeast lineages, and the differential retention of duplicates in the superorders Ostariophysi and Acanthopterygii following the Fish-Specific Genome Duplication (FSGD) may contribute to different evolutionary trajectories (Garcia de la Serrana et al. 2014). While these examples provide a tantalising hint that WGD may be a powerful evolutionary force, more studies are required across a range of polyploid systems to elucidate the evolutionary impacts of WGD.

Here, we used RAD data to investigate paralog retention and gene ontology differences in a family of neotropical catfish that have undergone at least two major large scale duplications (chapter 2 and chapter 3). The Callichthyidae are the biggest family in the order of the Siluriformes (Eschmeyer 2013) . The subfamily Corydoradinae is native to the neotropical Americas and consists of mostly herbivorous and insectivorous bottom-feeders. Previous work (see chapters 2 and 3) provided evidence for large-scale duplication events within the Corydoradinae catfish. Some species may have returned to a functionally diploid state, while mtDNA lineage 9 species appear to have undergone a more recent duplication event and show signs of functional polyploidy. Data indicate that the first major duplication

event occurred between lineages 1 and 2. As *A. poecilius* (lineage 2) has a similar genome size compared with *C. fowleri* (lineage 1) and SNP ploidy ratios indicate functional diploid for both species, it appears most likely that *A. poecilius* is a paleopolyploid. Potential additional duplication events may have occurred between lineages 3-8, though these are not detectable in our data sets.

Particularly in older polyploids where bivalent formation has been restored, duplicated loci no longer co-segregate and may diverge as independent paralogs (Wolfe 2001)**.** Here, we aim to identify whether tentative duplication events previously identified have led to differences in the genic makeup between lineages that could have impacted the subsequent evolution of the Corydoradinae. Specifically, we aim to use our RAD sequencing data set (described in detail in chapter 3) to address the following questions:

1. Species that have undergone large scale duplication events in the past may possess more genes (through functional divergence of paralogs over time), i.e. should possess more assembled contigs that blast as protein-coding genes.
2. Species that have undergone large-scale duplication events contain more paralogs, i.e. contain more contigs that are highly similar to one another as identified by Blastx.
3. Certain gene ontology groups have been preferentially retained after duplication events, i.e. certain gene ontology groups are enriched in comparison with pre-duplication species.

4.2 Methods

Data

RAD sequencing data were cleaned and de-multiplexed as described in chapter 3. Reads of individual samples were combined for each species and used to assemble the RAD data into contigs. In brief, we sequenced two sets of RAD libraries on Illumina HiSeq 2000, with each lineage of the Corydoradinae represented by at least one species with 2 individuals. Assemblies for all species were created using velvet and 'velvetoptimiser' (Gladman & Seeman 2012; Zerbino & Birney 2008) in two consecutive runs (as described in more detail in chapter 3). Coverage was normalized to assist assembly. In a first run, reverse reads were assembled individually to form stable contigs out of overlapping, sheared reads. In the second sequencing run, reverse read quality dropped below Q10 after 70 bases and these sequences were trimmed. This led to problems in the assembly, likely due to forward and reverse reads failing to overlap.

Raw reads for each sample were then mapped back to the assembled contigs using BWA-mem (Li 2013), and filtered to include only those contigs to which both forward and reverse reads of a same pair mapped. Contigs to which a minimum of ten read pairs mapped were analysed using Blastx as part of the NCBI blast+ suite (Camacho et al. 2009) to identify gene candidates.

Blastx Analysis and Paralog Identification

Blastx hits were filtered using the following criteria: Minimum length of 200 (to avoid several partial matches within a contig), an e value of $< 10^3$ (i.e. the chance of a false hit is smaller than 1 in 1000 for the given database) and a minimum of 30% similarity. For each species, filtered contigs were then blasted against themselves using Blastn in order to identify potentially paralogous sequences. Within species, Blastn hits also were filtered by length to avoid partial and false matches. Shorter reverse reads for samples of the second sequencing run produced significantly shorter reads (Analysis of Variance, $F=39.55$ $P<0.001$), which impaired the number of contigs assembled and may impair paralog-detection.

Contigs with a Blastx-hit fulfilling the above criteria were considered **putative genes**. Putative genes that had blast hits (minimum length of 200 bases, eval $<10^3$) to other putative genes in the same sample were considered **putative paralogs**.

Quantifying Haplotype Diversity of Blastx-Hits

Paralog identification was based on the contigs assembled in Velvet. Due to the nature of our short reads, and evidence for large scale duplication events within the Corydoradinae, many contigs have been identified as paralogous in previous analyses (see chapter 3 for more detail). In brief, paired-reads are too short to distinguish between highly similar regions in the genome, so that many assembled contigs do not represent a single unique region in the genome but a mixture of highly similar multi-copy gene families as well as recently duplicated genes. Thus, in order to test whether paralog and gene numbers may be underestimated, in this analysis we quantified haplotype number only for contigs with a Blastx hit (putative genes). As haplotype quantification is complicated by reads mapping to consecutive stretches of DNA that do not fully overlap (and would thus be identified as distinct haplotypes), we do not have haplotype information for every identified Blastx hit. Where the information was available, we identified all Blastx hits that contained more than 3 hits.

Gene Ontology Analysis

Blastx data for each sample was imported into Blast2GO (Conesa et al. 2005) for mapping and subsequent annotation using the default settings. To determine whether different species differ in their Gene Ontology (GO), we conducted two-tailed Enrichment Tests/Fisher's exact test integrated in Blast2GO using the FDR (False Discovery Rate) correction factor and a p-value cutoff of 0.05. As we were interested in GO differences within the sub-family that could be related to WGD events or gene duplications, we compared all species against *C. fowleri* (lineage 1).

As our Gene-data set is based on RAD data, our paralog and general annotation list is incomplete. The enrichment analysis as implemented in Blast2GO compares the proportion of sequences with the GO-term under consideration in the test sample against the proportion in the reference sample (Blüthgen et al. 2005).

Differences in GO terms across species and different categories were made more comparable by summarizing different Gene Ontology IDs into GO-slim categories using the AgBase GoSlim Viewer and using the Generic GO-slim set (McCarthy et al. 2006). GoSlim sets are subsets of GO-categories across all three ontologies and are useful for summarizing data and are maintained as part of the Gene Ontology database (Harris et al. 2004).

4.3 Results

Gene and Paralog Analysis

Data indicate a clear difference in the number of putative genes (Blastx-hits) obtained between sequencing runs (table 11). For sequencing run 1, hits obtained vary from 5189 to 7486 whereas for sequencing run 2, hits vary from 3286 to 3761. Putative paralogs range from 25 to 169 in sequencing run 2, whereas for sequencing run 1 between 537 and 2009 putative paralogs were identified**.** As the Blastx analysis was based on the assembled contigs, we tested for differences between the sequencing runs more thoroughly to establish whether the difference is a sequencing artefact or reflects a biological difference. Sequencing run 2 produced significantly fewer contigs than sequencing run 1(figure 12a**)** (Anova, F=20.01. df=1, p-value=0.00155). In addition, the number of contigs was significantly correlated not only with the number of Blastx-hits obtained (Pearson's correlation, cor. coefficient=0.9422, df=9, p-value<0.001) but also with the number of paralogs **(**Pearson's correlation, cor. coefficient=0.94437, df=9, p-value <0.001) (figures 12b and 12c**).** Thus, the drop in Blastx hits and the low number of paralogs appear to be the result of a quality drop in the second sequencing run.

Table 11. Blastx hits as well as putative paralogs identified using Blastn.

Lineage	Species	SeqRun	C-value	Blastx-Hits	Putative Paralogs	% Putative Paralogs
1	*C. fowleri*	*1*	0.65	5231	537	10.27
2	*A. poecilius*	*2*	0.76	3626	25	0.69
3	*S. kronei*	*2*	0.8	3761	44	1.17
4	*C. pygmaeus*	*2*	2.68	4636	169	3.65
5	*C. elegans*	*2*	2.24	3286	77	2.34
6	*C. nattereri*	*2*	1.79	2480	29	1.17
7	*C. aeneus*	*2*	1.84	3491	55	1.58
8	*C. imitator*	*1*	2.3	7486	2009	26.84
9	*C. metae*	*1*	4.15	5189	1714	33.03
9	*C. araguaiaensis*	*1*	4.36	5261	1541	29.29

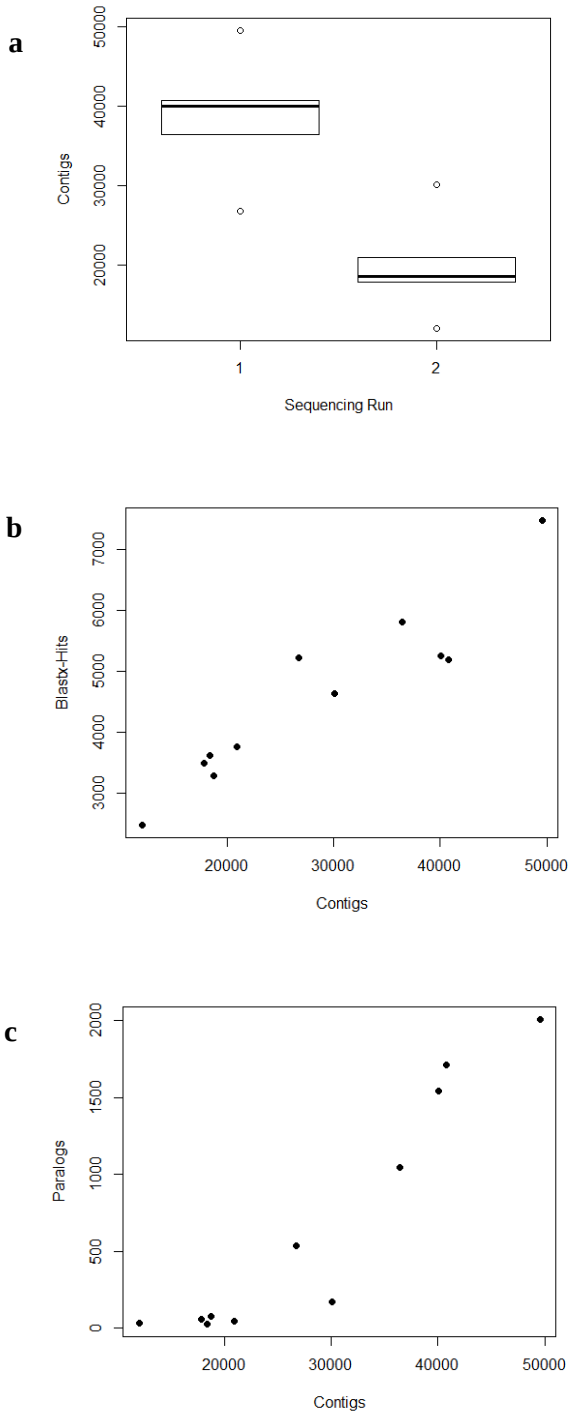


Figure 13. Examination of the potential effect of sequencing run on number of contigs and Blastx-hits recovered. a) Contig number retrieved for samples displayed by sequencing run. b) Paralogs identified against contigs assembled for each sample. c) Blastx-hits retrieved against contig number assembled for each sample.

The data for sequencing run 1 allows a direct comparison of a diploid species (*C. fowleri*) and species from lineages 8 and 9 which have undergone WGDs. There is a marked increase in the percentage of paralogs across these lineages - from 10.27% in *C. fowleri* (lineage 1) to 26-33% for *C. imitator, C. metae* and *C. araguaiaensis* (lineages 8 and 9). *C. metae* and *C. araguaiaensis* (lineage 9) have roughly the same number of Blastx hits as *C. fowleri* (lineage 1), even though previous analysis indicates that lineage 9 has undergone at least two additional rounds of large scale duplications compared to *C. fowleri*. The largest number of putative genes is found in *C. imitator* (lineage 8), a species that did not appear polyploid in previous analyses.

In order to test whether *C. metae* and *C. araguaiaensis* have a smaller number of putative genes because of the mis-assembly of multiple paralogous regions into single contigs, we identified the number of putative genes with multiple (more than 3) haplotypes where possible. As haplotype identification was impaired by partially overlapping reads, we restricted the haplotype analysis to shorter contigs. Table 12 outlines the percentage of Blastx hits for which we have haplotype information, as well as how many of these contain multiple haplotypes. In *C. fowleri*, 5% of all examined Blastx-hits contain multiple haplotypes, whereas from lineage 2 this shifts to above 20%. Generally, in lineages 2 - 8, between 12 and 25% of examined Blastx hits contain multiple haplotypes. This variation could in part be the result of the variation in the number of Blastx hits with data available. A shift is observed in lineage 9 with ~40% in *C. araguaiaensis* and ~30% in *C. metae* containing multiple haplotypes. However, due to restrictions in the haplotype analysis described earlier, we have only haplotype information for roughly 25-38% of putative genes. Thus, the number of putative genes, particularly for lineage 9, is likely to underestimate the true gene number, with many recent duplicates as well as closely related gene families mis-assembled into one contig.

Table 12. Percentage of Blastx-hits for which haplotype information was available, as well as the haplotype data for these. Only listed are the number of Blastx-hits with more than 3 haplotypes.

Lineage	Species	Sample	Percentage of Blastx-hits with haplotype information	Blastx-hits with > 3 haplotypes	Percentage of Blastx-hits with information with > 3 haplotypes
1	*C. fowleri*	Fow1	49.85	141	5.41
1	*C. fowleri*	Fow2	39.81	136	6.53
2	*A. poecilius*	Apo1	37.29	293	21.67
2	*A. poecilius*	Apo2	37.62	343	25.15
3	*S. kronei*	Skro1	40.20	243	16.07
3	*S. kronei*	Skro2	43.72	323	20.12
4	*C. pygmaeus*	Pyg1	24.14	235	21.00
4	*C. pygmaeus*	Pyg2	22.82	270	25.52
5	*C. elegans*	Eleg1	58.55	328	17.05
5	*C. elegans*	Eleg2	58.64	360	18.68
6	*C. nattereri*	Nat1	64.40	205	12.84
6	*C. nattereri*	Nat2	66.29	210	12.77
7	*C. aeneus*	Aen1	45.43	226	14.25
7	*C. aeneus*	Aen2	46.03	216	13.44
8	*C. imitator*	Imi1	42.41	352	11.09
8	*C. imitator*	Imi2	39.45	438	14.83
9	*C. metae*	Me1	24.61	553	43.30
9	*C. metae*	Me2	31.22	659	40.68
9	*C. araguaiaensis*	Ara1	36.11	603	31.74
9	*C. araguaiaensis*	Ara2	38.07	687	34.30

Gene Ontology Differences between Species

All species differ significantly from *C. fowleri* in several GO categories. *C. imitator* (lineage 8) and *C. metae* (lineage 9) show the largest differences in Gene Ontology, with *C. metae* showing significant differences in more than 120 individual GO categories, in most of which it appears underrepresented (table 13 and supporting information). When summarizing the differences as Go-slim categories across the three major gene ontology categories (Cellular Components, Molecular Functions and Biological Processes), trends across species become more apparent (see table 14). As the GO-slim terms are representative of a subset of different unique GO categories, species may appear over-represented as well as under-represented for the same GO-slim category. For Cellular Components, (particularly for the GO slim terms cellular_component, cell, and intracellular), several species from lineages 2, 8 and 9 are underrepresented and only *C. metae* is overrepresented in comparison to *C. fowleri* in specific categories. For molecular functions, this trend is the opposite, with GO-slim categories molecular_function and nucleotidyltransferase activity over-represented in several species, and other terms only underrepresented in *C. metae*. For biological processes, several species are overrepresented in DNA metabolic process, as well as biosynthetic process. Whereas a large number of species are overrepresented in the category biological process, some of these are also listed as underrepresented, again, due to differences in individual GO-categories summarized in the GO-Slim categories.

In summary, more species are under-represented for GO categories in Cellular Components, whereas more GO categories appear enriched in Molecular Function and Biological Processes.

Table 13. Blastx-hits, annotated genes as well as significant differences in GO-categories between all samples and *C. fowleri*. Differences are also listed as GO-Slim categories.

Lineage	Species	Blastx-hits	Annotated Genes	GO vs. *C. fowleri*	GO-SLIM Differences
1	*C. fowleri*	*5231*	*3228*	-	-
2	*A. poecilius*	*3626*	*2032*	9	6
3	*S. kronei*	*3761*	*2368*	3	4
4	*C. pygmaeus*	*4636*	*2787*	3	4
5	*C. elegans*	*3286*	*2014*	5	6
6	*C. nattereri*	*2480*	*1613*	4	5
7	*C. aeneus*	*3491*	*2194*	3	4
8	*C. imitator*	*7486*	*3918*	22	9
9	*C. metae*	*5189*	*2845*	128	27
9	*C. araguaiaensis*	*5261*	*2905*	9	7

Table 14. All Go-Slim categories and the species in which these are significantly over-represented or under-represented. As GO-Slim categories contain a subset of other GO categories, some species appear both over- and under-represented.

Go-Slim ID	Go-Slim Descriptions	Over-represented in	Under-represented in
Cellular Components			
Slim id: GO:0005575	cellular_component	*C. metae*	*A. poecilius, C. imitator, C. araguaiaensis, C. metae*
Slim id: GO:0005622	intracellular	*C. metae*	*C. imitator, C. metae*
Slim id: GO:0005623	Cell	*C. metae*	*A. poecilius, C. imitator, C. araguaiaensis, C. metae*
Slim id: GO:0005737	cytoplasm	*C. metae*	*C. metae*
Slim id: GO:0005794	Golgi apparatus		*C. metae*
Slim id: GO:0005886	plasma membrane		*C. metae*
Slim id: GO:0043226	organelle		*C. metae*
Slim id: GO:0043234	protein complex	*C. metae*	
Molecular Functions			
Slim id: GO:0003674	molecular_function	*A. poecilius, S. kronei, C. pygmaeus, C. elegans, C. nattereri, C. aeneus, C. imitator,C. araguaiaensis,C. metae*	*C. metae*
Slim id: GO:0004871	signal transducer activity	*C. metae*	
Slim id: GO:0016301	kinase activity		*C. metae*
Slim id: GO:0016779	Nucleotidyltransferase activity	*A. poecilius, C. elegans, C. araguaiaensis, C. metae*	
Slim id: GO:0043167	ion binding		*C. metae*
Biological Processes			
Slim id: GO:0006259	DNA metabolic process	*A. poecilius, S. kronei, C. pygmaeus, C. elegans, C. nattereri, C. aeneus, C. imitator,C. araguaiaensis, C. metae*	
Slim id:	cellular protein modification		*C. metae*

GO:0006464	process		
Slim id: GO:0006629	lipid metabolic process		*C. metae*
Slim id: GO:0006810	transport		*C. metae*
Slim id: GO:0007165	signal transduction	*C. metae*	*C. imitator, C. metae*
Slim id: GO:0008150	biological_process	*A. poecilius, S. kronei, C. pygmaeus, C. elegans, C. nattereri, C. aeneus, C. imitator,C. araguaiaensis, C. metae*	*A. poecilius, C. imitator, C. araguaiaensis, C. metae*
Slim id: GO:0009058	biosynthetic process	*A. poecilius, C. elegans, C. nattereri, C. araguaiaensis, C. metae*	*C. metae*
Slim id: GO:0009790	embryo development		*C. metae*
Slim id: GO:0030154	cell differentiation		*C. metae*
Slim id: GO:0032196	transposition	*A. poecilius, S. kronei, C. pygmaeus, C. elegans, C. nattereri, C. aeneus, C. imitator,C. araguaiaensis, C. metae*	
Slim id: GO:0034641	cellular nitrogen compound metabolic process	*C. metae, C. imitator*	*C. metae*
Slim id: GO:0040007	Growth		*C. metae*
Slim id: GO:0044281	small molecule metabolic process		*C. metae*
Slim id: GO:0048856	anatomical structure development		*C. metae*

4.4 Discussion

In this study, we show that species which have undergone large scale duplication events possess a higher percentage of multi-copy putative genes (based on Blastx-hits), as well as a higher percentage of putative paralogs. The Gene Ontology analysis revealed differences across species, with cellular components generally under-represented compared to *C. fowleri* and biological processes and molecular functions over-represented. Unfortunately, a quality drop in the reverse reads of the second sequencing run affected the assembly and resulted in significantly fewer contigs which impaired analyses in this chapter. Thus, the main focus of this discussion is on species from sequencing run 1, namely *C. fowleri* (lineage 1), *C. imitator* (lineage 8), *C. metae* and *C. araguaiaensis* (lineage 9).

Gene and Paralog Numbers

Very early on it was realized that fish contain more genes than other vertebrates as a result of ancient WGD (Wittbrodt et al. 1998). Despite the loss of the majority of duplicate genes, the retention of Ohnologs (paralogs through WGD) appears to be a common feature across many taxa, and gene retention is strongly correlated with duplication mode (e.g. Hakes et al. 2007; Freeling 2009). Thus, if the large scale duplication events detected in chapter 3 are the result of WGD events (as would be suggested by the *Hox* data set in chapter 2), we would expect a detectable increase in the number of genes as well as paralogs in a specific subset of categories.

We identified a clear increase in paralogs, with lineages 8-9 displaying roughly 30% paralog- content compared to roughly 6% in lineage 1. In teleost fish studied to date, between 17-21% of paralogs from the ancient FSGD have been retained (Garcia de la Serrana et al. 2014), but as our analysis is based on a RAD data set and we do not have the full complement of genes, it is perhaps not surprising that we only identify 5% putative paralogs in the basal *C. fowleri* . Particularly for lineage 9, putative paralogs are likely highly underestimated due to the large number of multi-copy genes detected. Bearing this in mind, other model polyploid plant species such as maize (*Zea mays*), soybean (*Glycine max*) and wheat (*Triticum aestivum*) were estimated to contain between 32 and 37% of paralogs based on EST sequences (Blanc & Wolfe 2004). Both *C. metae* and *C. araguaiaensis* have similar paralog abundances to these known polyploids and are therefore consistent with being polyploids.

In terms of the number of putative genes identified, *C. fowleri* (lineage 1), *C. metae* and *C. araguaiaensis* (both lineage 9) contain roughly the same number of putative genes. Surprisingly, *C. imitator* (lineage 8) contained some two thousand more putative genes than the other species. Therefore, despite a large fraction of duplicated putative paralogs in *C. araguaiaensis* and *C. metae*, there appears to be no overall increase in putative gene number. As teleosts possess far more genes than tetrapods as the result of the FSGD (Wittbrodt et al. 1998), with current estimates of retained copies averaging 15% (Braasch & Postlethwait 2012), we would expect that several rounds of WGD in the Corydoradinae would also lead to an increase in putative gene numbers detected. When considering that a minimum of two large duplication events have occurred, *C. araguaiaensis* and *C. metae* have proportionally fewer putative genes than we would expect. This could be an artefact of the RAD data, with additional putative genes in lineage 9 remaining undetected. In addition, total gene numbers are underestimated, with the haplotype analysis indicating that roughly 30% of putative genes in lineage 9 represent multi-copy genes. Thus, the additional gene copies we would be expecting may be mis-assembled into one contig and thus present as multi-copy putative genes in this analysis.

Another plausible explanation is genome shrinkage or genome downsizing. Genome downsizing is well documented in plants and describes the phenomenon that in many polyploid organisms, the relative haploid genome size is smaller than would be expected (Leitch & Bennett 2004). It has been proposed that genome downsizing may facilitate restoring meiosis to its diploid behaviour and thus lead to cytological rediplodization and can result from unequal recombination, often induced by LTR elements (Eilam et al. 2010). Genome downsizing and reciprocal loss of gene duplicates may facilitate speciation and contribute to diversification rates. For instance reciprocal gene loss shortly after WGD events has been shown to be a major factor in the evolution and speciation of yeast (Scannell et al. 2006; Scannell et al. 2007; Morel et al. 2015). Sémon & Wolfe (2007) estimated that thousands of genes underwent reciprocal gene loss when comparing the genomes of *Tetraodon* and *Danio rerio* and suggest that reciprocal gene loss could have also played a role in teleost diversification.

We identified the largest number of putative genes in *C. imitator* (lineage 8). *C. imitator* also contains many fewer multi-copy putative genes when compared with lineage 9 species, albeit the number of paralogs is very similar. The high gene and paralog number in

comparison with lineage 9 is peculiar, as lineage 9 has undergone an additional duplication event when compared to lineage 8. Such apparent disparity could reflect a lineage-specific difference in gene retention. Alternatively, the larger number of contigs identified as a gene alongside the lower number of multi-copy gene contigs may indicate that the overall divergence of paralogous genes in *C. imitator* is higher than in lineage 9, which could be indicative of an allopolyploidy event.

When quantifying the number of haplotypes for putative genes (Blastx-hits) where available, the pattern is congruent with the analysis of all contigs in chapter 3: A marked shift in multi-copy Blastx-hits between lineage 1 and lineage 2 is detected, as well as in lineage 9. We were unable to reliably quantify selection type and strength of our paralogs and multi-copy genes directly due to short reads and the partial hits of putative genes. However, inferences can be made based on their similarity. As explained in chapter 3, in a perfect assembly contigs should represent one chromosomal location with a maximum of 2 alleles in a heterozygous diploid. In *C. metae* roughly 40% of identified putative genes/Blastx-hits contain a minimum of 3 haplotypes, indicating that several genes belonging to different genomic locations have been mis-assembled into one contig because they were too similar for the assembler to be able to distinguish between them. This is based on very short reads of a given gene hit, but indicates that at least part of these genes retained high similarity, which in turn could indicate that a large proportion of these genes may remain under purifying selection. Alternatively, the high similarity could indicate a very recent duplication event, or that similarity across duplicates is maintained through the formation of multivalents at meiosis. Such assertions would be in line with the expectation that both *C. metae* and *C. araguaiaensis* are likely functional polyploids as discussed in chapter 3.

Gene Ontology Differences

It is now well documented that the mode of gene duplication affects gene retention (Freeling 2009; Wang et al. 2012). While the results for individual GO categories should be interpreted with caution, an overall trend is apparent: a reduction in the category ‘Cellular Components’, and an increase in the categories ‘Biological Processes’ and ‘Molecular Function’ in species compared to the basal diploid *C. fowleri*. The trend is congruent with a study on *Tetraodon* (Brunet et al. 2006) which found a similar overall pattern after the FSGD

event. Despite the overall loss of data for sequencing run 2, species part of this sequencing run still display the same trend in GO retention or underrepresentation as samples from sequencing run 1. That this trend is apparent from lineage 2 onwards (e.g. *A. poecilius*) further supports the notion that the Corydoradinae underwent at least two rounds of Whole Genome Duplication with one occurring between lineage 1 and lineage 2.

Curiously, the GO-Category 'DNA Metabolism' appears enriched in all species when compared to *C. fowleri* whereas this category has been reported to be underrepresented after WGD in *Arabidopsis* (Freeling 2009). Furthermore, our data shows an opposite pattern to that detected in Compositae plants where structural and cellular components were enriched and transcription factors, molecular functions and metabolic processes were underrepresented (Barker et al. 2008). However, as Barker et al. (2008) note, while similar retention often occurs within lineages, these patterns tend to differ between higher taxonomic groups. Similarly to the Corydoradinae, metabolic genes have been retained preferentially in the moss after a WGD event *Physcomitrella patens* (Rensing et al. 2007).

Corydoras metae shows a much higher number of differences in GO than any other species, with many of these differences contrasting with overall observed trends. For example, *C. metae* is under-represented in many Biological Processes or Molecular Functions even though it matches over-representation in categories such as 'DNA Metabolism' or 'transposition', which are over-represented across species. These differences could in part be an artefact of the incomplete nature of the RAD-data, though it is curious that we see such a high degree of differentiation only in *C. metae*. In the absence of further chromosomal and genome data, we can only speculate on the cause. For instance, Warren et al (2014) were able to distinguish between WGD derived paralogs and tandem events based on whether paralogs were located on the same chromosome. This revealed that Atlantic salmon (*Salmo salar)* underwent increased local duplications after the salmon-specific WGD. There did not appear to be a difference in GO between paralogs created via WGD or tandem events. However, tandem duplications have been shown to have opposite retention patterns to WGD events in plants (Freeling 2009).

Conclusion

While there are limitations in inferring genic information and gene ontology from a RAD data, we nevertheless were able to identify overall trends in the Corydoradinae that appear similar to those identified in *Tetraodon*. We identified an increase in paralogs and

multi-copy genes in line with expectations for species that have undergone major duplication events. Thus, this analysis provides further evidence that the uncovered large scale duplication events were indeed polyploidization events. To address hypothesis 1, we could not find direct evidence for an increase in gene number after duplication events, though it is acknowledged that the quality of RAD data was not optimal. We were however able to show that in accordance with hypotheses 2 and 3, there is a detectable increase in paralogs after WGD events as well as a clear difference in Gene Ontology.

4.5 Supporting Information

Table 15. List of Gene Ontology terms over-represented in Corydoradinae species compared with *Corydoras fowleri*.

GO Term	Description	Over-represented in
GO:0034061	DNA polymerase activity	*A. poecilius, C. araguaiaensis, C. aeneus , C. metae*
GO:0003964	RNA-directed DNA polymerase activity	*A. poecilius, C. elegans, C. araguaiaensis, C. aeneus , C. metae*
GO:0006278	RNA-dependent DNA replication	*A. poecilius, C. elegans, C. nattereri, C. araguaiaensis, C. aeneus , C. metae*
GO:0006313	transposition, DNA-mediated	*A. poecilius, S. kronei, C. pygmaeus, C. elegans, C. nattereri, C. aeneus, C. imitator, C. araguaiaensis, C. aeneus, C. metae*
GO:0004803	transposase activity	*A. poecilius, S. kronei, C. pygmaeus, C. elegans, C. nattereri, C. aeneus, C. imitator, C. araguaiaensis, C. aeneus, C. metae*
GO:0032196	Transposition	*A. poecilius, S. kronei, C. pygmaeus, C. elegans. C. nattereri, C. aeneus, C. imitator, C. araguaiaensis, C. aeneus , C. metae*
GO:0090304	nucleic acid metabolic process	*C. imitator, C. metae*
GO:0006807	nitrogen compound metabolic process	*C. imitator, C. metae*
GO:0015074	DNA integration	*C. imitator, C. metae*
GO:0006259	DNA metabolic process	*C. imitator, C. metae*
GO:0006310	DNA recombination	*C. imitator, C. metae*
GO:0006139	nucleobase-containing compound metabolic process	*C. metae*
GO:0034641	cellular nitrogen compound metabolic process	*C. metae*
GO:0006725	cellular aromatic compound metabolic process	*C. metae*
GO:0046483	heterocycle metabolic process	*C. metae*
GO:0070098	chemokine-mediated signalling pathway	*C. metae*
GO:1901360	organic cyclic compound metabolic process	*C. metae*
GO:0004950	chemokine receptor activity	*C. metae*
GO:0001637	G-protein coupled chemo-attractant receptor activity	*C. metae*
GO:0003676	nucleic acid binding	*C. metae*
GO:0005960	glycine cleavage complex	*C. metae*

Table 16. Gene Ontology terms under-represented in Corydoradinae species compared with *Corydoras fowleri*.

GO Term	Description	Under-represented
GO:0044699	single-organism process	*A. poecilius, C. imitator, C. araguaiaensis, C. aeneus , C. metae*
GO:0044464	cell part	*A. poecilius, C. imitator, C. araguaiaensis, C. aeneus , C. metae*
GO:0005623	Cell	*A. poecilius, C. imitator, C. araguaiaensis, C. aeneus, C. metae*
GO:0065007	biological regulation	*C. imitator, C. metae*
GO:0050794	regulation of cellular process	*C. imitator, C. metae*
GO:0050896	response to stimulus	*C. imitator, C. metae*
GO:0007154	cell communication	*C. imitator, C. metae*
GO:0051716	cellular response to stimulus	*C. imitator, C. metae*
GO:0050789	regulation of biological process	*C. imitator, C. metae*
GO:0023052	Signaling	*C. imitator, C. metae*
GO:0044700	single organism signaling	*C. imitator, C. metae*
GO:0007165	signal transduction	*C. imitator, C. metae*
GO:0005622	Intracellular	*C. imitator, C. metae*
GO:0016020	Membrane	*C. imitator, C. metae*
GO:0009653	anatomical structure morphogenesis	*C. metae*
GO:0048731	system development	*C. metae*
GO:0007399	nervous system development	*C. metae*
GO:0048513	organ development	*C. metae*
GO:0048856	anatomical structure development	*C. metae*
GO:0040008	regulation of growth	*C. metae*
GO:0040007	Growth	*C. metae*
GO:0044271	cellular nitrogen compound biosynthetic process	*C. metae*
GO:0034654	nucleobase-containing compound biosynthetic process	*C. metae*
GO:2001141	regulation of RNA biosynthetic process	*C. metae*
GO:0006355	regulation of transcription, DNA-dependent	*C. metae*
GO:0032774	RNA biosynthetic process	*C. metae*
GO:0006351	transcription, DNA-dependent	*C. metae*
GO:0051252	regulation of RNA metabolic process	*C. metae*
GO:0016070	RNA metabolic process	*C. metae*
GO:0019438	aromatic compound biosynthetic process	*C. metae*
GO:0018130	heterocycle biosynthetic process	*C. metae*
GO:1901362	organic cyclic compound biosynthetic process	*C. metae*
GO:0009889	regulation of biosynthetic process	*C. metae*

GO:0031326	regulation of cellular biosynthetic process	*C. metae*
GO:2000112	regulation of cellular macromolecule biosynthetic process	*C. metae*
GO:0010556	regulation of macromolecule biosynthetic process	*C. metae*
GO:0008203	cholesterol metabolic process	*C. metae*
GO:0016125	sterol metabolic process	*C. metae*
GO:0006468	protein phosphorylation	*C. metae*
GO:0016043	cellular component organization	*C. metae*
GO:0071840	cellular component organization or biogenesis	*C. metae*
GO:0044267	cellular protein metabolic process	*C. metae*
GO:0010467	gene expression	*C. metae*
GO:0051179	localization	*C. metae*
GO:0043412	macromolecule modification	*C. metae*
GO:0007275	multicellular organismal development	*C. metae*
GO:0006793	phosphorus metabolic process	*C. metae*
GO:0016310	phosphorylation	*C. metae*
GO:0060255	regulation of macromolecule metabolic process	*C. metae*
GO:0044707	single-multicellular organism process	*C. metae*
GO:0044767	single-organism developmental process	*C. metae*
GO:0044765	single-organism transport	*C. metae*
GO:0032502	developmental process	*C. metae*
GO:0051234	establishment of localization	*C. metae*
GO:0032501	multicellular organismal process	*C. metae*
GO:0006796	phosphate-containing compound metabolic process	*C. metae*
GO:0019538	protein metabolic process	*C. metae*
GO:0036211	protein modification process	*C. metae*
GO:0065008	regulation of biological quality	*C. metae*
GO:0031323	regulation of cellular metabolic process	*C. metae*
GO:0010468	regulation of gene expression	*C. metae*
GO:0019222	regulation of metabolic process	*C. metae*
GO:0080090	regulation of primary metabolic process	*C. metae*
GO:0044763	single-organism cellular process	*C. metae*
GO:0006810	transport	*C. metae*
GO:0006464	cellular protein modification process	*C. metae*
GO:0005524	ATP binding	*C. metae*
GO:0035639	purine ribonucleoside triphosphate binding	*C. metae*
GO:0043168	anion binding	*C. metae*

GO:0043167	ion binding	*C. metae*
GO:0004672	protein kinase activity	*C. metae*
GO:0004674	protein serine/threonine kinase activity	*C. metae*
GO:0016301	kinase activity	*C. metae*
GO:0030554	adenyl nucleotide binding	*C. metae*
GO:0032559	adenyl ribonucleotide binding	*C. metae*
GO:0097367	carbohydrate derivative binding	*C. metae*
GO:0001882	nucleoside binding	*C. metae*
GO:1901265	nucleoside phosphate binding	*C. metae*
GO:0000166	nucleotide binding	*C. metae*
GO:0016773	phosphotransferase activity, alcohol group as acceptor	*C. metae*
GO:0005515	protein binding	*C. metae*
GO:0001883	purine nucleoside binding	*C. metae*
GO:0017076	purine nucleotide binding	*C. metae*
GO:0032550	purine ribonucleoside binding	*C. metae*
GO:0032555	purine ribonucleotide binding	*C. metae*
GO:0032549	ribonucleoside binding	*C. metae*
GO:0032553	ribonucleotide binding	*C. metae*
GO:0036094	small molecule binding	*C. metae*
GO:0043231	intracellular membrane-bounded organelle	*C. metae*
GO:0043229	intracellular organelle	*C. metae*
GO:0043227	membrane-bounded organelle	*C. metae*
GO:0043226	organelle	*C. metae*
GO:0044459	plasma membrane part	*C. metae*
GO:0005886	plasma membrane	*C. metae*
GO:0005794	Golgi apparatus	*C. metae*
GO:0044444	cytoplasmic part	*C. metae*
GO:0005737	cytoplasm	*C. metae*
GO:0044424	intracellular part	*C. metae*
GO:0030424	axon	*C. metae*
GO:0071944	cell periphery	*C. metae*
GO:0044463	cell projection part	*C. metae*
GO:0016021	integral to membrane	*C. metae*
GO:0031224	intrinsic to membrane	*C. metae*
GO:0044425	membrane part	*C. metae*
GO:0030154	cell differentiation	*C. metae*

GO:0009790	embryo development	*C. metae*
GO:0009888	tissue development	*C. metae*

Chapter 5 - Transposable Elements Expansion and Whole Genome Duplication drive genome size variation in Corydoradinae catfish

5.1 Introduction

Transposable Elements (TEs) are repetitive DNA sequences that have the ability to move within a genome. There are two main classes of TEs that can be distinguished by their mode of transposition. Class I elements, also known as retrotransposons, transpose via a RNA intermediate using a 'copy-and-paste' mechanism. Class II elements, or DNA transposons, use a DNA intermediate (or in some cases no intermediate) via a 'cut-and-paste' mechanism (Finnegan 1989; Wicker et al. 2007). Transposable Elements were first discovered by Barbara McClintock who identified a TE insertion in a mutant allele involved in kernel pigmentation patterns in maize (*Zea mays*) (McClintock 1951) . TEs are an almost universal presence in genomes and have been shown to be a significant force in genome size evolution in both animals and plants (e.g. Feschotte et al. 2002; Chénais et al. 2012; Lee & Kim 2014). For example, 85% of the genome is composed of TEs in maize (*Zea mays*) (Schnable et al. 2009). Other examples where TE proliferation has led to an increase in genome size include the rice genus (*Oryza sp.*) (Zuccolo et al. 2007), cotton (*Gossypium sp.*) (Hawkins et al. 2006) and salamanders (Sun et al. 2012). In humans, TEs make up roughly 45% of the genome (International Human Genome Consortium 2001).

TEs have often been considered as purely "selfish elements" due to their predominantly deleterious nature (Doolittle & Sapienza 1980; Orgel & Crick 1980). Among their effects, TE insertions can interrupt or alter host gene functions, and lead to recombination between non-homologous chromosomes causing large scale genomic rearrangements and deletions (Montgomery et al. 1991; Finnegan 1992; Mieczkowski et al. 2006; Brawand et al. 2014; Lee 2015). To avoid such deleterious insertions, TE activity is usually tightly controlled via epigenetic silencing/methylation and specific RNAi-pathways in the host genome. Previously silenced TEs for instance, are transcribed and then broken into siRNAs (small interfering RNAs) which can act as a homology-based sensor, targeting similar

elements for methylation in the genome (e.g. Aravin et al. 2007; Kim & Zilberman 2014). It is not well understood how new TE-elements (entering the genome through horizontal transfer) become silenced, though several mechanisms have been proposed (reviewed in Fultz et al. 2015). Furthermore, environmental stress, hybridization and particularly polyploidy have been shown to interrupt silencing mechanisms, leading to the re-activation and subsequent proliferation of TE elements (Capy et al. 2000; Oliver et al. 2013; Casacuberta & González 2013; Fultz et al. 2015).

The view that TEs have only deleterious effects is now changing with numerous examples of TEs impacting on host gene function and structure (e.g. Kidwell & Lisch 2000). Indeed, many TEs have been domesticated by the host genome and incorporated into protein-coding genes (Volff 2006; Feschotte & Pritham 2007; Chénais et al. 2012). Moreover, TE activity may have evolutionary impacts and facilitate adaptation to environmental change. For example, a TE insertion appears to have led to resistance against insecticides such as DDT in *Drosophila* (Chung et al. 2007). The capacity of TEs to create genetic variation through changes in gene expression, alternative splicing, exon shuffling or genomic rearrangements has also been proposed as a mechanism through which invasive species overcome a lack of genetic variation resulting from small founder populations (Schrader et al. 2014).

Corydoradinae catfish are a highly speciose family of neo-tropical catfish that display enormous variation in genome size, with C-values ranging from 0.6 to 4.4 pg of DNA. In chapters 2 and 3, evidence was uncovered that suggests that the Corydoradinae have undergone several rounds of WGD in their evolutionary history, which could potentially explain the enormous variation in genome size. The aim of this study was to determine whether other factors have also contributed to variation in genome size, with a particular emphasis on TEs. Our objectives were

1) to quantify the abundance of repetitive elements across species
2) to quantify the relative importance of both WGD events and repetitive elements in the genome size expansion of the Corydoradinae.

5.2 Methods

Data Acquisition

In this chapter, we further analysed the raw RAD data from Chapter 3 with a particular focus on TE element abundance. To summarize, we sequenced two RAD library sets for ten species across the Corydoradinae as well as one outgroup on a HiSeq2000 Illumina Sequencer (250bp paired-end). Each mitochondrial lineage was represented by at least one species, with 2 individual samples per species. As the analysis in this chapter was based on raw reads and not assembled contigs, the effect of the quality drop in the second sequencing run (described in chapter 3 and 4) should be minimal: while reverse reads of affected species may be shorter, this would not affect the comparison of overall percentage of bases identified as repetitive between runs.

Identification of TE elements

We identified and quantified repeats and TEs in the RAD data of each species using Repeatmasker using default settings and specifying "teleost species" as the target group (Smit et al. 2013-2015.). We used Repbase**,** a database for repetitive elements, as the reference database to search against (Jurka et al. 2005). In order to avoid potential PCR bias, we de-replicated all raw reads using Usearch (Edgar 2010) with the 'derep_fulllength' option prior to the Repeatmasker Analysis.

Repeat-Types and Comparative Mapping

In addition to identifying the main super-families of Transposable Elements, we further analysed the Repeatmasker output using Excel and custom made scripts to identify copy numbers of Repeat-Classes and Repeat-Families. We compared these for presence/absence across lineages. We also used comparative mapping approaches to assess similarity of masked sequences across lineages. For this we used the unmasked *C. fowleri* (lineage 1) assembly as a baseline and mapped reads from all species back to this baseline using the mapping algorithm BWA-mem (Li 2013). We repeated this with the masked *C.*

fowleri assembly as a reference. This would allow us to assess similarity across species based on transposable element content.

Drivers of Genome Size

In order to determine the main drivers behind the genome size increase in the Corydoradinae, we aimed to quantify the relative importance of TEs versus the number of WGD events as a potential contributor to genome size. We thus combined evidence from chapters 2 and 3 and categorized species by the number of potential WGDs implicated in previous chapters (table 17). For lineages 2, 5 and 6, different species were investigated in chapters 2 and chapters 3 respectively with highly similar genome sizes. Note that strictly speaking, we did not find any evidence for an additional duplication event in either lineage 5 or lineage 8, though as both lineage 4, 6 and 7 appear to have undergone an additional event and genome sizes are very similar, we assume the most parsimonious explanation is that lineage 5 and 8 share a second genome duplication event for the purposes of this analysis. An ANCOVA model was performed in R studio (RStudio 2012) with C-values (based on species from chapter 3) as the response, WGD events as a categorical variable and TE-abundance (in percent) as a continuous variable.

Table 17. A star indicates a shared WGD event implied by analyses in specified chapters. + denotes evidence for an additional event in this species/lineage. Results from both chapters were summarized as a factor in WGD events and used for downstream analysis.

Lineage	HOX	RAD	C- value	WGD Events
0			1.58pg	Nill
1			0.51-0.65pg	Nill
2	*	*	0.76pg	One
3			0.8pg	One
4	+		2.2-2.68pg	Two
5			2.02-2.24pg	(Two)
6	+	+	1.62-1.79pg	Two
7		+	1.84pg	Two
8			1.96-2.3pg	(Two)
9	*	*	3.2-4.36pg	Three

5.3 Results

Genome Size and Transposable Element Content

The masked data revealed large differences in repetitive elements among species and lineages. Lineages with larger genome sizes have a higher abundance of repetitive elements (figure 14). There is a slight increase in TE abundance between lineages 1 and 3, with an abrupt increase in lineages 4-6 with more than 20% TE content. More than half the data in lineage 7 (*C. aeneus*) comprises repetitive elements - a five times increase when compared to lineage 1**.** There is a slight drop in TE abundance in lineage 8 (*C. imitator*), with the highest abundance of TE elements in any lineage found in lineage 9.

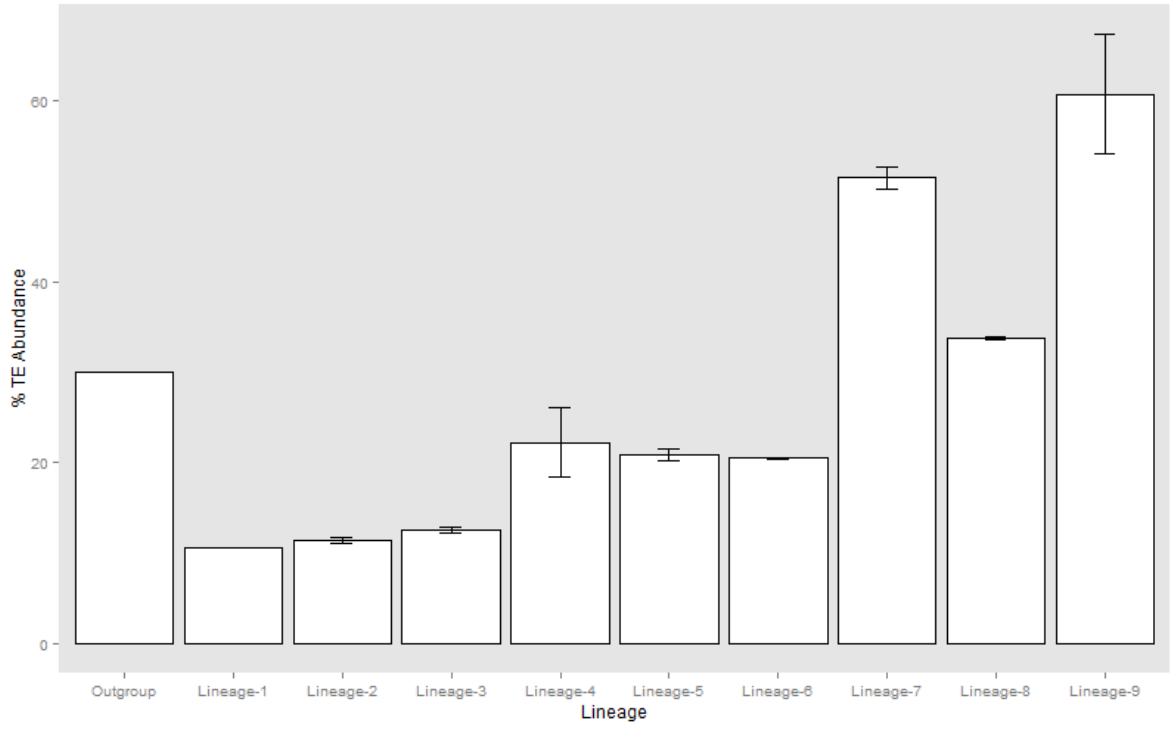


Figure 14. Repeat content in percent identified by Repeatmasker displayed by lineage. The outgroup is represented by only one individual sample for *Megalechis*, there are thus no error bars for this individual.

Contributors of TE-Expansion

The abundance of different Retroelements and DNA-Transposons identified by Repeatmasker in each of the species and individuals is detailed in table 18. Most TE families remain relatively stable across the samples investigated with the exception of DNA-Transposons, specifically the superfamily TC1-IS630-Pogo (table 18). The percentage variation in each across species is shown in figure 15. We further investigated copy number variation in the Repeat-Classes/Families belonging to the above listed superfamilies of Retroelements and DNA-Elements. The ten largest Repeat-Classes/Families identified are displayed in figure 16 with the dominant family being TcMar-TC1, belonging to TC1-IS630-Pogo.

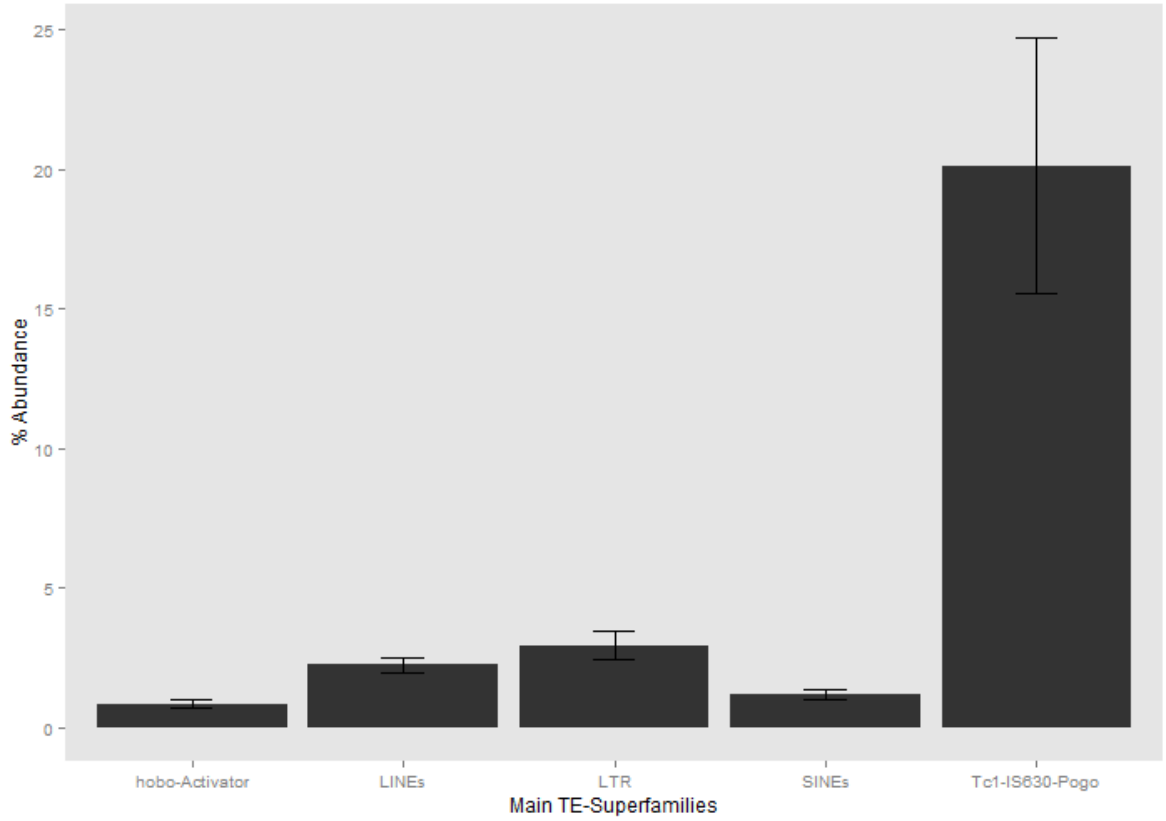


Figure 15. Variation in abundance (percentage of bases) across species for the main TE-families identified.

Table 18. Mega-Basepairs (MBP) and % identified as one of the 5 major superfamilies of TEs for each species.

		Retroelements						DNA-Transposons					
Lineage	Species	SINEs		LINEs		LTR elements		hobo-Activator		TC1-IS630-Pogo		Total Masked	
		MBp	%	MBp	%	MBp	%	MBp	%	MBp	%	Mbp	%
0	*Megalechis*	1.53	1.81	13.69	1.62	0.29	0.35	0.22	0.26	17.49	20.68	25.28	42.07
1	*C. fowleri-1*	2.45	0.91	87.80	3.26	4.80	1.78	0.54	0.20	1.66	0.62	28.32	10.50
1	*C. fowleri-2*	2.66	0.94	89.94	3.19	4.54	1.61	0.58	0.21	2.49	0.88	29.61	10.49
2	*A. poecilius-1*	0.50	1.43	5.13	1.45	0.44	1.24	0.08	0.22	1.23	3.49	4.11	11.66
2	*A. poecilius-2*	0.79	1.28	8.52	1.37	0.70	1.13	0.14	0.22	2.01	3.23	6.81	10.96
3	*S. kronei-1*	0.29	0.84	6.74	1.94	0.45	1.29	0.08	0.24	1.71	4.91	4.25	12.23
3	*S. kronei-2*	0.34	0.92	7.04	1.92	0.51	1.38	0.11	0.29	1.81	4.94	4.72	12.90
4	*C. pygmaeus-1*	1.48	2.03	29.39	4.04	1.89	2.59	0.34	0.47	4.03	5.53	13.37	18.36
4	*C. pygmaeus-2*	3.57	4.15	34.22	3.98	4.73	5.50	0.48	0.55	5.26	6.12	22.34	25.98
5	*C. elegans-1*	0.43	0.96	20.72	4.62	0.75	1.68	0.16	0.36	4.41	9.85	9.61	21.45
5	*C. elegans-2*	0.43	0.96	19.87	4.42	0.76	1.70	0.16	0.36	4.01	8.92	9.08	20.23
6	*C. nattereri-1*	0.28	0.80	7.75	2.25	1.94	5.62	0.53	1.54	2.36	6.83	7.02	20.34
6	*C. nattereri-2*	0.37	0.92	9.53	2.38	2.15	5.37	0.55	1.37	2.70	6.72	8.23	20.52
7	*C. aeneus-1*	0.31	0.42	5.18	0.70	0.92	1.26	0.28	0.38	35.46	48.20	38.77	52.69
7	*C. aeneus-2*	0.37	0.47	5.67	0.73	1.10	1.41	0.46	0.59	35.13	45.05	39.21	50.28
8	*C. imitator-1*	2.49	1.64	38.36	2.53	3.28	2.16	1.85	1.22	34.10	22.46	50.97	33.58
8	*C. imitator-2*	3.58	2.04	44.80	2.55	4.63	2.64	2.16	1.23	37.65	21.45	59.44	33.86
9	*C. metae-1*	2.05	0.39	36.98	0.70	14.41	2.72	3.37	1.91	371.70	70.22	396.18	74.85
9	*C. metae-2*	3.08	0.47	61.55	0.94	22.46	3.42	11.47	1.75	395.55	60.27	440.11	67.06
9	*C. araguaiaensis-1*	2.81	0.59	56.84	1.19	39.25	8.19	8.29	1.73	176.18	36.78	233.52	48.75
9	*C. araguaiaensis-2*	2.64	0.56	55.37	1.17	38.95	8.26	10.29	2.18	166.20	35.22	224.88	47.66

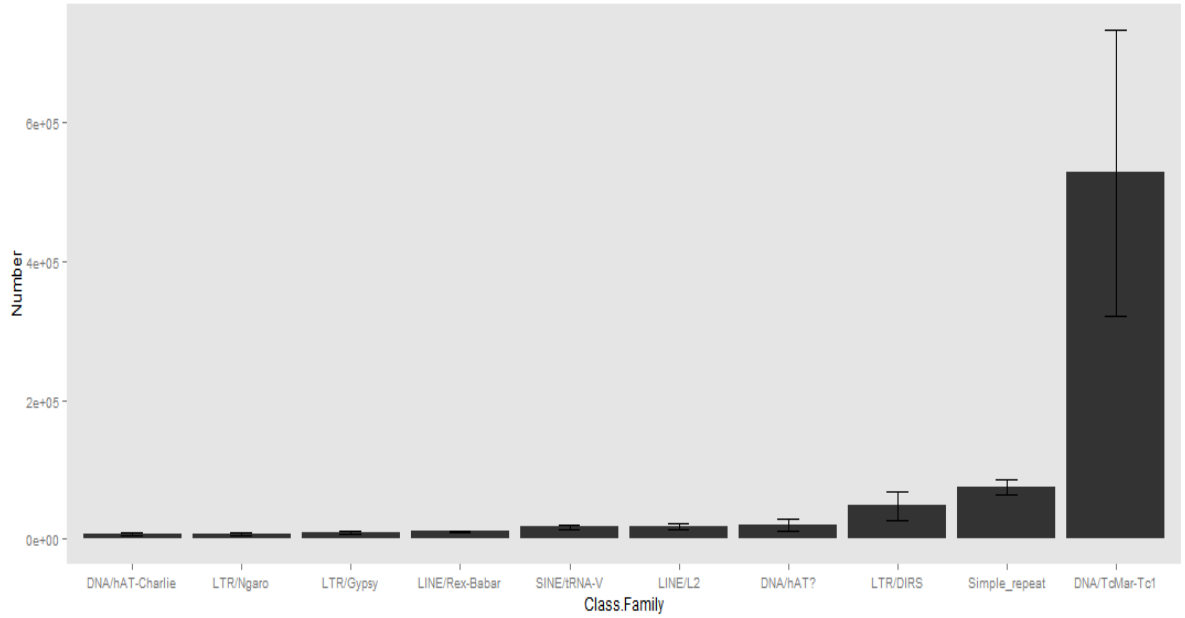


Figure 16. Read Number of the ten most abundant TE-Families across species. All Families were identified in all 21 Samples.

The comparative mapping approach (figure 17) revealed that the elements that have expanded in lineages 7-9 (*C. aeneus*, *C. imitator*, *C. metae* and *C. araguaiaensis*) are also present/ highly similar to those present in lineage 1. For example, roughly half of all *C. metae* sequence data map successfully to the unmasked *C. fowleri* assembly. This similarity entirely disappears after masking, indicating that the mapping success was driven by repetitive elements identified by Repeatmasker.

A. poecilius (lineage 2) also displays a large discrepancy between masked and unmasked mapping similarity, with a large standard error for the unmasked data. This is likely the result of large variation in the number of sequencing reads for *A. poecilius* (see figure 9 in chapter 3).

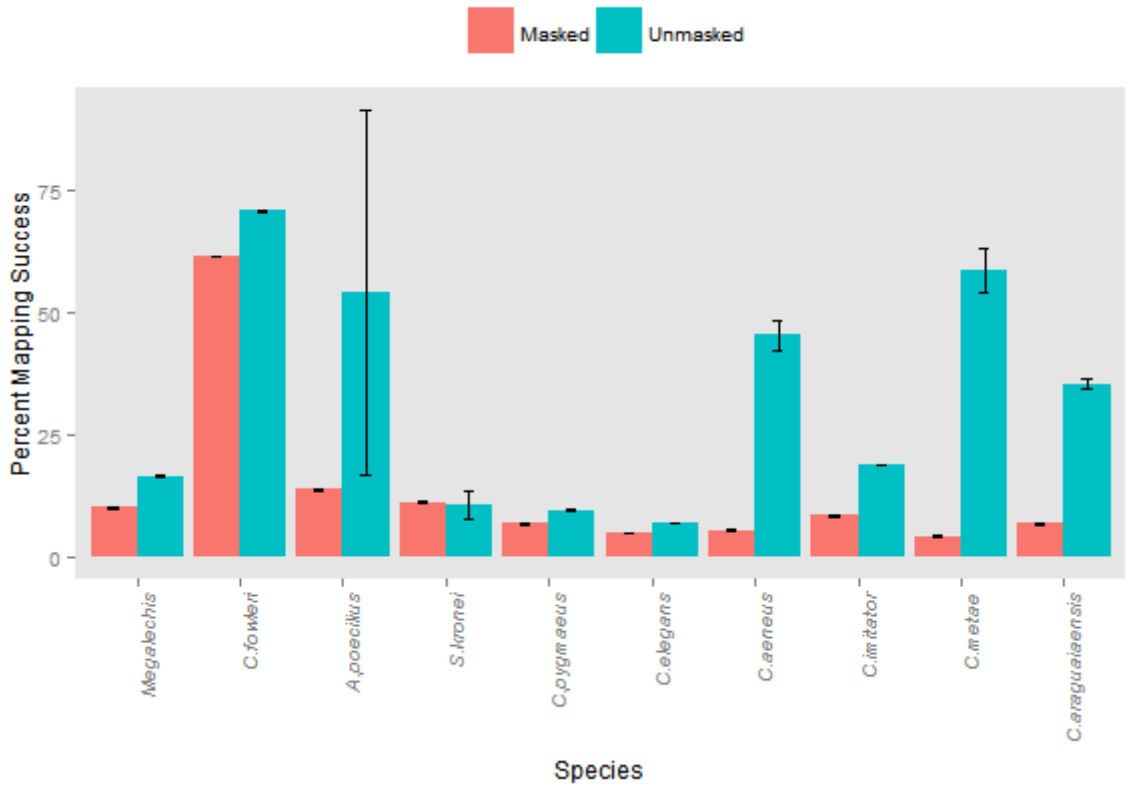


Figure 17. All species were mapped to contigs assembled for *Corydoras fowleri*. Percentage of reads that mapped successfully is displayed for all species for when contigs were masked or unmasked.

Whole Genome Duplication vs. Transposable Elements

The boxplots (figure 18) display the variation in TE-content and C-values across groups that have undergone between zero and three WGD events (table 17). Both show a similar trend; species that have undergone additional WGD events appear to contain more TEs and also appear to have higher genome sizes. The variation in TEs appears to be far more variable after 2 or 3 WGD events (figure 18.a). In addition, genome size (measured as C-values) and TE-abundance are significantly correlated as visualized in figure 19 (Pearson's product-moment correlation, $t=5.8726$, $df=19$, $p<0.001$).

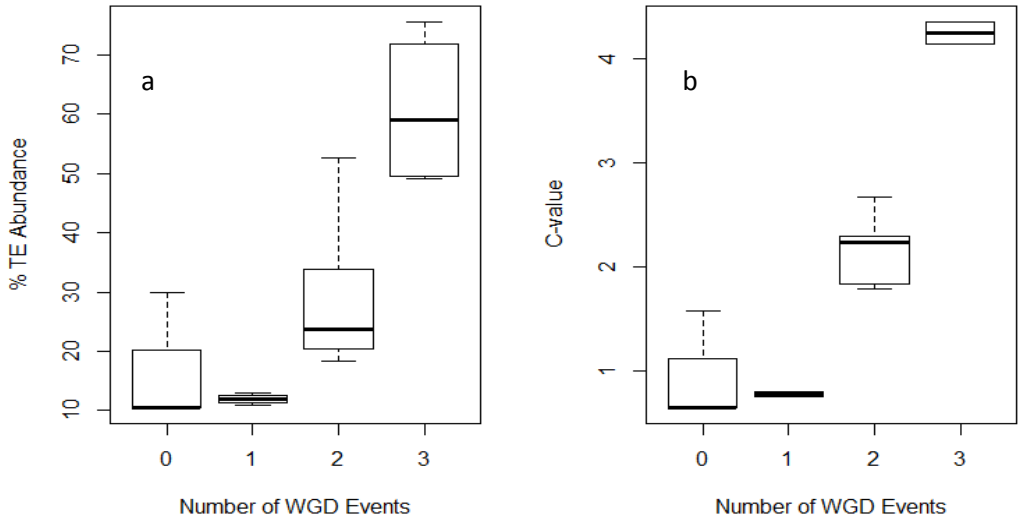


Figure 18. Variation in TE content and genome size relative to the number of identified WGD events. a) Variation in % TE abundance is plotted against the number of WGD events. b) Genome Size (C-Value) is plotted against number of WGD events.

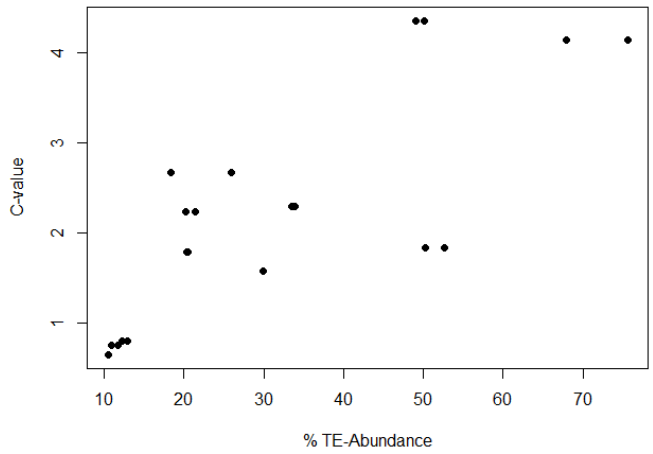


Figure 19. TE-abundance in percent is plotted against genome size.

Ancova analysis

The Ancova model was used to partition the relative importance of TE content and number of WGD events as drivers of genome size (table 19). The model explained roughly 96% of the variation in C-values with a p-value of <0.001. Both TE-abundance and WGD Events emerge as significant variables. WGD-2 and WGD-3 have t-values of 6.173 and 6.168 respectively and appear to have the largest and most significant impact on genome size, whereas WGD-One is non-significant. The more recent a WGD event occurred, the more significant its impact on genome size was found to be. WGD- 2 and WGD-3 interact significantly with TE-abundance which indicates that these WGD events and the TE proliferation may be linked.

Table 19. Results obtained from the Ancova analysis listing details of the model. Note that the intercept in this case represents the first categorical variable WGD-Nill.

Coefficients	**Estimate**	**Std.Error**	**t-value**	**Pr(>|t|)**	
(Intercept)	0.1466	0.30988	0.473	0.6440	
TE-Abundance (%)	0.04796	0.01609	2.982	0.0106	*
WGD-One	-0.19647	0.38011	-0.517	0.6139	
WGD-Two	2.32561	0.37672	6.173	0.000305	***
WGD-Three	4.64931	0.75372	6.168	0.000139	***
WGD-One:TE	0.02166	0.01900	1.140	0.2747	
WGD-Two:TE	-0.05688	0.01956	-2.908	0.0122	*
WGD-Three:TE	-0.05961	0.01745	-3.416	0.0046	**

Significance denoted as: * p<0.05; ** p<0.01; *** P<0.001

Residual standard error: 0.2547 on 13 degrees of freedom

Multiple R-squared: 0.973, Adjusted R-squared: 0.9584

F-statistic: 66.8 on 7 and 13 DF, p-value: 3.542e-09

5.4 Discussion

In this study, we quantified the repetitive element content across Corydoradinae lineages using a RAD data set and identified that TE proliferation was driven by one DNA-Transposon family, namely Tc1-like elements. Subsequently, we used a modelling approach to identify the main drivers of genome expansion in the group. While the major driver of the genome size increase was the number of WGD events, TE proliferation was also a significant contributor.

Polyploidy and TE proliferation

Multiple WGD events and TE proliferation have led to a more than 6 fold increase in genome size in the Corydoradinae. The first WGD event detected between lineages 1 and 3 has had no significant effect on genome size, though this is most likely due to extensive re-diploidization as discussed in previous chapters. Genome size increase from lineage 4 onwards appears to be driven mainly by WGD events and TE elements. The overall effect of the TE abundance, however, is weaker than the effect of WGD events. However, the interaction between TE abundance and number of WGD events is also significant, with TE abundance increasing with every WGD event which suggests that TE proliferation occurred as a consequence of WGD events.

Numerous examples exist where both WGD events and TE proliferation have contributed to genome size variation. Two prominent examples in the plant kingdom include rice (*Oryza species*) as well as maize (*Zea mays*) (Zuccolo et al. 2007; Schnable et al. 2009), and both WGD and TEs have been implicated in the evolution of the hugely diverse angiosperms (Oliver & Greene 2009). Due to their deleterious mutagenic potential, host genomes usually silence TEs epigenetically. However, polyploidy and hybridization may interrupt the suppression mechanisms employed by the host genome, allowing TEs to proliferate in the genome (Oliver & Greene 2009). Thus, the 2nd and 3rd WGD events in the Corydoradinae may have interrupted such control mechanisms and led to the increase of Tc1-like DNA transposons. Furthermore, it has been suggested that WGDs may buffer some of the potentially deleterious consequences of TE insertion (Cañestro & Albalat 2012; Oliver et al. 2013). In addition, an increased presence of TEs in a polyploid genome may aid the formation of bivalents through increased homolog divergence, leading to a more stable meiosis and

increased gamete fertility (Oliver & Greene 2009). However, in cotton as well as in wheat, the TE proliferation appears to be independent of the duplication or hybridization events, indicating that TE re-activation and proliferation is not always a consequence of WGD events (Charles et al. 2008; Hu et al. 2010). We would need longer TE-sequences to identify the age of the TE proliferations in the Corydoradinae, which may yield additional insights into the cause. Another plausible mechanism that could have led to re-activation of TE-elements is environmental stress: TE mobilization has been demonstrated in response to for instance temperature changes, UV light stress, pathogens and infections (Chénais et al. 2012; Oliver et al. 2013; Casacuberta & González 2013). Such associations may be a response of the genome to these stressors through increased mutations and thus adaptive resources. Examples of adaptations through TE activity include the development of insecticide resistance in *Drosophila melanogaster* (Chung et al. 2007), gene expression in response to light signalling in *Arabidopsis* (Lin et al. 2007), adaptation to latitudes in soybean (Liu et al. 2008; Kanazawa et al. 2009), as well as adaptations in *Drosophila* in response to climate variation (González et al. 2010; Kim et al. 2014).

Proliferation of Tc1-like elements

The Tc1-superfamily belongs to the subclass of DNA transposons and is widespread across a wide range of organisms. Tc1-elements have been identified in fungi, plants, ciliates and animals, particularly fish and amphibians (Radice et al. 1994; Robertson 1995; Nandi et al. 2007; Pocwierz-Kotus et al. 2007). Tc1-elements are usually around 1.7kb long and contain a gene encoding for a transposase enzyme (that enables the movement of the element) flanked by inverted repeats (Radice et al. 1994; Robertson 1995). Most copies present in teleosts and vertebrates in general contain frameshift mutations within their transposase genes rendering them inactive, though active Tc1 elements have been identified in the Japanese medaka (Kawakami et al. 2000; Nandi et al. 2007). There is also evidence for transcription of Tc1-like transposons in salmonids and catfish, though this may not mean that transposition events are occurring (Krasnov et al. 2005; Nandi et al. 2007). Initial analysis of the Channel Catfish genome (*Ictalurus punctatus*) showed that Tc1-elements are the most prominent and widespread TE-elements in this species, making up roughly 4-5% of the genome. Tc1-elements also appear far more evenly spread across the genome, as opposed to other elements that were more clustered (Nandi et al. 2007; Jiang et al. 2011). Such trends could mean that our data set based on RAD-tags (which are spread across the genome) is biased towards

picking up Tc1-like elements, and may result in an underestimate of other TE-families that may be more clustered.

Despite such ambiguity, it is clear that Tc1-elements have expanded in the Corydoradinae and successfully escaped suppression in the higher genome size lineages. As previously discussed, genomic shock through polyploidy or hybridization, or environmental stressors, can lead to the interruption of methylation and small RNA suppression machinery. The increase in Tc1-elements appears to occur in lineage 2, lineage 5, lineage 7 and lineage 9, the same points for which we found evidence for WGD events. *Megalechis* on the other hand also has a high Tc1-element content and does not appear to be polyploid nor paleopolyploid. Alternatively, new Tc1-elements (or elements sufficiently diverged from those already recognized by siRNAs in the genome) may enter the host genome through hybridization or through parasites and spread before the genome is able to control for their presence. Little is known to date about how similar a TE-element has to be to already suppressed elements for siRNA to successfully recognize it and activate methylation responses (Fultz et al. 2015). However, in our case, the elements that have expanded appear similar enough to those already present in *C. fowleri* for a mapping algorithm to recognize them. Without further data, it is thus not possible to identify the trigger for the expansion.

Conclusion and Future Perspectives

While genome size variation in the Corydoradinae appears largely driven by WGD events, an increase in TE abundance particularly from lineage 7 onwards has also had a significant effect. Such expansion is driven by Tc1-like elements, which also appear to have been incorporated in genes across the Corydoradinae and could be of adaptive value. More data are necessary to study the proliferation of TEs in more detail. It is unlikely that high quality whole genome sequences will become available soon, particularly for the paleopolyploid and polyploid species. More traditional methods like BAC cloning and PCR sequencing of complete elements or alternatively nanopore /Pacific Biosciences sequencing technologies (Metzker 2010; Schneider & Dekker 2012) across species could allow for the construction of phylogenies which could help to identify potential hybridization events.

5.5 Supporting Information

Table 20. Outline of assembled contigs and their repetitive element content. Note that many TE-elements have been assembled into one contig, likely due to their high similarity, leading to a lower percentage of masked bases in comparison with the read data. A higher diversity in higher lineages indicates higher repetitive element diversity.

Species	Contigs	Bases	GC level	Masked Bases	% Masked Bases
C. fowleri	26743	10877571	42.85	1288023	11.84
A. poecilius	18374	5497558	43.06	475045	8.64
S. kronei	20866	6921133	42.24	652134	9.42
C. pygmaeus	30051	10506449	40.61	1487750	14.16
C. elegans	18674	5486920	42.69	606783	11.06
C. nattereri	11972	3245528	44.25	396128	12.21
C. aeneus	17787	5125424	43.37	576735	11.25
C. imitator	49552	21894466	40.94	3399133	15.53
C. metae	40763	16433566	40.88	3127615	19.03
C. araguaiaensis	40026	16326542	41.4	2816067	17.25

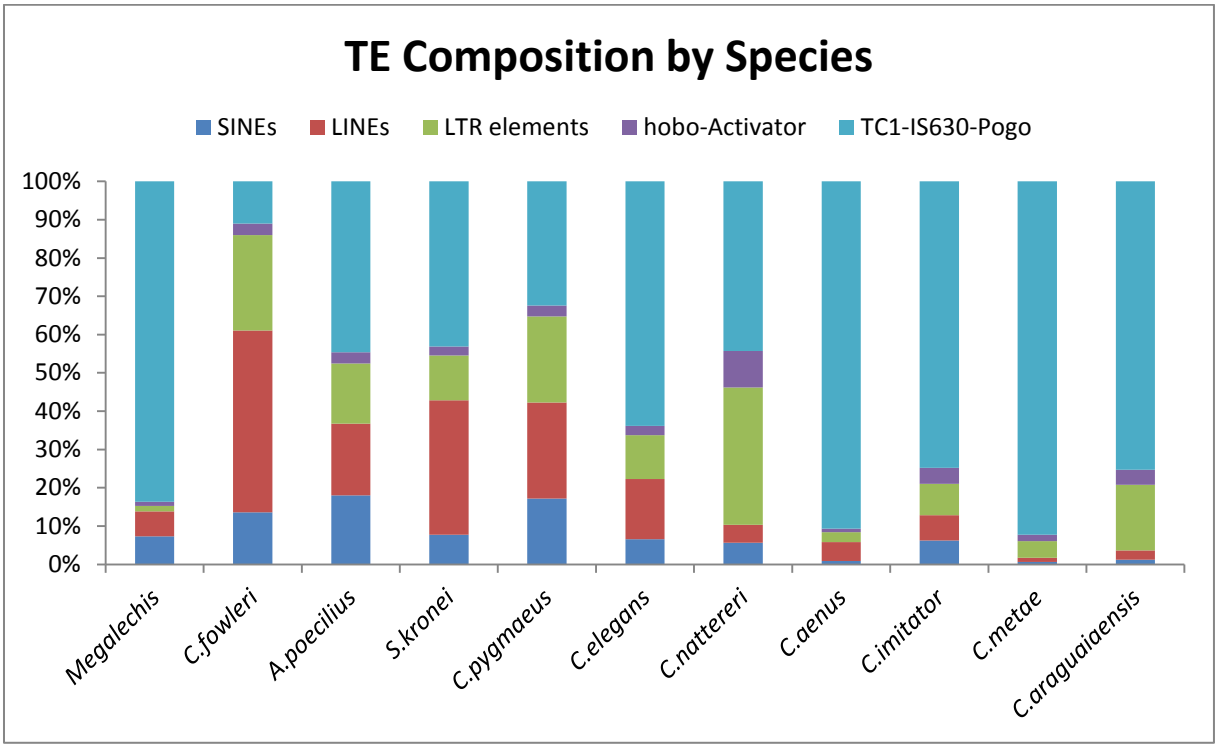


Figure 20. Composition of Transposable Elements across species.

Chapter 6 - Discussion, Conclusion and Future Perspectives

6.1 Overview of Results

The aim of this thesis was to identify the underlying mechanism driving extensive genome size variation in the Corydoradinae catfish. We accomplished this using a *Hox* gene as a WGD marker as well by studying a RAD sequencing data set for representative species from across all of the mtDNA lineages. Together, these data provide compelling evidence for multiple rounds of Whole Genome Duplication (WGD), as well as for DNA-transposon expansion. Both the WGD events identified, as well as the Transposable Element (TE) content, significantly contribute to genome size in the Corydoradinae.

The aim of chapter 2 was to use the *HoxA13a* gene as a marker for WGD events by quantifying copy number per species. In this chapter, phylogenetic analysis of *HoxA13a* sequences across species allowed us to identify likely points of WGD events, with one or likely several shared WGD events occurring between lineages 1 and 3, as well as independent duplication events in lineages 4, 6 and 9.

Chapter 3 also aimed to identify WGD events as well as functional ploidy status across the Corydoradinae using a RAD sequencing data set. In this chapter, we assembled paired reads into contigs. Instead of attempting to remove paralogous regions, we quantified them by identifying the number of haplotypes at any given contig. This allowed us to detect significant shifts of multi-copy contigs in the Corydoradinae, confirming duplication events we previously identified in chapter 2. Excitingly, SNP read ratios also indicate that several species in lineage 9 and perhaps lineage 6-7 could be functionally polyploid. In addition, we found that a lineage 6 species (*C. nattereri*), appears more closely related to lineage 9 species on the basis of the RAD data.

In chapter 4, protein-coding regions in the RAD-data set were identified, and potentially paralogous regions quantified. A significant increase in putative paralogs was

detected. Additionally, and despite working with an incomplete data set, Gene Ontology analysis showed similar enrichment patterns to those described previously after the fish specific whole genome duplication in teleosts. Thus, this chapter provides further evidence that WGD has played a major role in the Corydoradinae.

Repetitive and Transposable Elements were quantified in chapter 5. A significant increase in Tc1-DNA Transposons was detected in lineage 7 and appears to have significantly contributed to genome size increase in this subfamily. Genome size variation can almost completely be explained by TE abundance when taken together with WGD events previously identified in chapter 2 and chapter 3.

Table 21. Summary of results including WGD events (based on chapter 2 and chapter 3), ploidy status (chapter 3) and % TE Content (chapter 5).

Lineage	WGD Events	Polyploid?	% TE Content
0	Nill		25.08
1	Nill		7.77
2	One		8.02
3	One		9.73
4	Two		18.59
5	Two?		17.51
6	Two	?	17.73
7	Two	?	50.37
8	Two?		30.82
9	Three	?	70.96
9	Three	?	48.21

6.2. Discussion of Results

Identification of WGD Events

In summary, our analyses confirm that several rounds of WGD events have taken place in the evolutionary history of the Corydoradinae, some of which are shared and some of which appear lineage or species specific events. Surprisingly, we originally predicted several rounds of WGD to have occurred much deeper in the phylogeny, and assumed that lineages 1 to 3 were entirely diploid. However, both the *Hox* data as well as the RAD analysis clearly show that a major WGD event occurred between lineages 1 and 2. As genome sizes are nearly identical between lineages 1 to 3, and SNP ratios (chapter 3) show very similar profiles to that of lineage 1, it appears that lineages 2 and 3 have re-diploidized. This is by no means a unique scenario. *Arabidopsis thaliana* was initially thought to be a diploid with a small genome size, but was found to be an ancient paleopolyploid when the genome was sequenced (The Arabidopsis Genome Initiative 2000)**.**

In the *Hox* data set, we considered clades formed of multiple species as an individual gene copy. Altogether, we have identified six of these paralogous copies, which all contain species representative of lineages 3 to 9, indicating up to three duplication events between lineages 1 and 3. A potential scenario is an initial duplication or hybridization event leading to triploidy, with two triploids forming a hexaploid species. This is purely speculative, though may explain why we see three diverged paralogous groups in the *HoxA13a*.

The *Hox* data and the RAD data set complement each other well, and both show WGD events between lineages 1 and 2 and within lineage 9, as well as on a further event in lineage 6. Further support for WGD events comes from the Gene Ontology, which found patterns strikingly similar to those identified in teleosts after the FSGD (Brunet et al. 2006). However, some duplication events were only picked up in the *Hox* data set, whereas lineage 6 and 7 duplications only became apparent in the SNP ratio analysis. These duplications may well be restricted to species analysed, with different representing lineage 6 in the *Hox* and RAD data sets, and warrant further lineage- wide investigations.

Ploidy level, ploidy mode and multivalent inheritance

Due to extensive gene loss and genome wide deletions after WGD events (Wolfe 2001), identifying WGD events by themselves does not provide information about the current ploidy level. However, the RAD data set allowed us to look at read count frequencies for bi-allelic SNPs, which has been successfully used to identify functional polyploids in the potato blight causing pathogen *Phytophthora infestans* (Yoshida et al. 2013) and more recently in *Arabidopsis arenosa* (Arnold et al. 2015).

This revealed that *C. araguaiaensis* and *C. metae* appear to be functionally polyploid, and potentially also *C. aeneus* and *C. nattereri*. These read frequencies are suggestive of multivalent formation or continuous recombination between homeologous chromosomes, one of the key diagnostics used to distinguish autopolyploids from allopolyploids (Parisod et al. 2010). Allopolyploids are more likely to form strict bivalents during meiosis (and if thus differentiated also likely to be assembled into different contigs), leading to diploid-like inheritance patterns and thus diploid like ½ SNP read ratios. Particularly in recent polyploids, which lineage 9 individuals may well be, multivalent formation is often seen as evidence for autopolyploidy (Gregory & Mable 2005). However, it is widely appreciated that there is a continuum between the doubling of identical genomes to the doubling of highly differentiated genomes, which can present as a mixture of disomic and tetrasomic inheritance patterns, with some homeologous chromosomes forming multivalents, and others forming strict bivalents (Stebbins 1971; Otto 2007). Thus, the SNP read ratios are not strictly evidence for strict autopolyploid origins, though at the same time it appears unlikely that *C. metae* and *C. araguaiaensis* are allopolyploids with highly differentiated parental species.

Despite our SNP read ratios indicating polyploidy, multivalent formation has not been reported in the Corydoradinae (personal communication Martin Taylor and Claudio Oliveira), which may seem counterintuitive. However, multivalent formation is not necessary to achieve the patterns of tetrasomic inheritance observed. For instance, in natural populations of the autotetraploid *Arabidopsis arenosa*, homeologs form bivalents, but randomly pair up and thus continue to display tetrasomic inheritance (Carvalho et al. 2010; Yant et al. 2013). This allows them to escape the deleterious effects of multivalent formation. It is plausible that a similar mechanism exists in many other species, including the Corydoradinae.

6.3 Future Work

This thesis has helped shed light on two mechanisms that contribute significantly to genome size variation within the diverse subfamily of the Corydoradinae. However, many questions remain to be answered. These initial findings may lay the foundation for future research, focusing not only on the genome evolution per se, but also on its phenotypic effects and the ecology of the Corydoradinae.

Whole Genome Duplication and Ploidy levels

Despite a clear signature of WGD events in the Corydoradinae, questions as to the exact ploidy status and origin remain. A reference genome would greatly aid further study of the Corydoradinae, and would enable comparative mapping approaches. In the absence of whole genome sequencing, a transcriptomic approach could perhaps also lead to greater insights into when these events have occurred, as well as how many and what kind of genes have been preferentially retained.

In order to successfully distinguish between autopolyploids, allopolyploids and segmental allopolyploids (particularly in lineage 9), sequencing of more species could greatly aid identification of potential within-lineage hybridization events. Controlled multi-generation inheritance studies for several markers, in conjunction with cytological mapping approaches, could also give further insights into the tetrasomic inheritance patterns discovered here.

We also identified a clear conflict between the mitochondrial based phylogeny and our RAD data set, with *C. nattereri* (a lineage 6 species) appearing to be more closely related to lineage 9 than lineages 7 and 8. The question is whether this holds for all species of lineage 6, which could indicate a hybridization event or allopolyploidy event leading to the formation of lineage 9 species, or whether this is an isolated species effect.

Transposable Elements

The short RAD reads prevented us from a phylogenetic analysis of TEs identified in the Corydoradinae, which could aid the identification of potential hybridization events. Full-length TE-sequences would also allow us to time these expansions more accurately, and to determine the age of individual elements and whether or not they remain active. TE activity can lead to chromosomal rearrangements, deletions and changes in epigenetics, thus potentially leading to reproductive isolation and increasing speciation rates (Hurst & Werren 2001; Volff 2005; Oliver et al. 2013). Intriguingly, previous research shows a significant increase in net diversification rates for lineages 7-9 (Alexandrou 2011), which could potentially be linked to TE activity. It has been speculated that TEs could have played a role in fish speciation and diversification, though this has yet to be shown empirically (Volff 2005) .

Adaptive Evolution of the Corydoradinae- Immunity and Colour Pattern Evolution

The Corydoradinae are also an ideal system to study the effect of WGD on immunity. Preliminary data on immunity indicated that species with a higher genome size may have a smaller number of parasites (unpublished). We now know that these communities consist of species of different ploidy level, or in the very least of species that have undergone a different number of independent WGD events. One of the potential advantages of polyploidy could be related to increased parasite resistance resulting from an increase in diversity of the major histocompatibility complex (MHC), a theory first suggested by Levin (1983)and more recently mathematically explored by Oswald and Nuismer (Oswald & Nuismer 2007). Oswald and Nuismer (2007) argue that initially, novel polyploid lineages are more resistant than their diploid progenitors, providing a plausible explanation for successful establishment of polyploid lineages. It would be extremely interesting to quantify differences in response to diseases and parasites, and whether there is a link between immunity and ploidy levels.

Nothing is known about the genetic basis of colour pattern variation and mimicry in the Corydoradinae. Intriguingly, the largest number of species as well as the largest number of mimics are found in lineages 8 and 9, which leads to the question whether WGD may have facilitated the evolution of colour pattern convergence or enabled these species to “copy” colour patterns of other lineages. The FSGD event for instance has greatly diversified

pigmentation genes in the teleosts and has likely led to the stunning diversity of pigmentation patterns in fish, with 30% of pigmentation genes preferentially retained in duplicate (Braasch et al. 2009).

In summary, the Corydoradinae make an ideal model system to study not only the complex genomic changes after WGD and polyploidy in vertebrates, but also offer opportunities to study the interactions between polyploidy and TE diversification, as well as the genomic basis of adaptive traits such as mimicry or an increase in immunity. Next generation sequencing technologies and ever decreasing costs will aid the investigation of this fascinating group of Corydoradinae catfish, a powerful system to study genomic and adaptive evolution. My hope is that this thesis has successfully built a foundation that allows testing of hypotheses in relation to adaptive evolution and genome evolution in the Corydoradinae.

Bibliography

Albert, J., Petry, P. & Reis, R.E., 2011. Major Biogeographic and Phylogenetic Patterns. In J. Albert & R. E. Reis, eds. *Historical Biogeography of Neotropical freshwater fishes*. Berkeley: University of California Press, pp. 21–57.

Albertin, W. & Marullo, P., 2012. Polyploidy in fungi: evolution after whole-genome duplication. *Proceedings Of The Royal Society B: Biological Sciences*, 279(1738), pp.2497–2509.

Alexandrou, M., 2011. *Mechanisms of Speciation and Coexistence in Corydoradinae Catfishes*. PhD Thesis Bangor University.

Alexandrou, M., Oliveira, C., Maillard, M., McGill, R. a R., Newton, J., Creer, S. & Taylor, M.I., 2011. Competition and phylogeny determine community structure in Müllerian co-mimics. *Nature*, 469(7328), pp.84–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21209663 [Accessed March 1, 2012].

Alexandrou, M. & Taylor, M.I., 2011. Evolution, Ecology and Taxonomy of the Corydoradinae Revisited. In I. A. M. Fuller & H.-G. Evers, eds. *Identifying Corydoradinae Catfish - Supplement 1*. Ian Fuller Enterprises, pp. 104–141.

Aljanabi, S.M. & Martinez, I., 1997. Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques. *Nucleic Acids Research*, 25(22), pp.4692–4693.

Allendorf, F.W., Bassham, S., Cresko, W. a., Limborg, M.T., Seeb, L.W. & Seeb, J.E., 2015. Effects of Crossovers Between Homeologs on Inheritance and Population Genomics in Polyploid-Derived Salmonid Fishes. *Journal Of Heredity*, 106(3), pp.1–11. Available at: http://jhered.oxfordjournals.org/cgi/doi/10.1093/jhered/esv015 [Accessed May 5, 2015].

Amores, A., Force, A., Yan, Y., Joly, L., Amemiya, C., Fritz, A., Ho, R.K., Langeland, J., Prince, V., Wang, Y., Westerfield, M., Ekker, M. & Postlethwait, J.H., 1998. Zebrafish *hox* Clusters and Vertebrate Genome Evolution. *Science*, 282(5394), pp.1711–1714. Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.282.5394.1711 [Accessed July 4, 2011].

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J.-M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., Gelpke, M.D.S., Roach, J., Oh, T., Ho, I.Y., Wong, M., Detter, C., Verhoef, F., Predki, P., Tay, A., Lucas, S., Richardson, P., Smith, S.F., Clark, M.S., Edwards, Y.J.K., Doggett, N., Zharkikh, A., Tavtigian, S. V, Pruss, D., Barnstead, M., Evans, C., Baden, H., Powell, J., Glusman, G., Rowen, L., Hood, L., Tan, Y.H., Elgar, G., Hawkins, T., Venkatesh, B., Rokhsar, D. & Brenner, S., 2002. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science (New York, N.Y.)*, 297(5585), pp.1301–1310.

Aravin, A. A, Hannon, G.J. & Brennecke, J., 2007. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science (New York, N.Y.)*, 318(5851), pp.761–764.

Arnold, B., Bomblies, K. & Wakeley, J., 2012. Extending coalescent theory to autotetraploids. *Genetics*, 192(1), pp.195–204. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3430536&tool=pmcentrez&rendertype=abstract [Accessed May 14, 2015].

Arnold, B., Corbett-Detig, R.B., Hartl, D. & Bomblies, K., 2013. RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology*, 22(11), pp.3179–3190. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23551379 [Accessed May 13, 2015].

Arnold, B., Kim, S., Bomblies, K. & Biology, E., 2015. Single geographic origin of a widespread autotetraploid *Arabidopsis arenosa* lineage followed by interploidy admixture. *Molecular Biology And Evolution*.

Baird, N. A, Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z. A, Selker, E.U., Cresko, W. a & Johnson, E. A, 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS One*, 3(10), p.e3376. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2557064&tool=pmcentrez&rendertype=abstract [Accessed January 30, 2013].

Barker, M.S., Kane, N.C., Matvienko, M., Kozik, A., Michelmore, R.W., Knapp, S.J. & Rieseberg, L.H., 2008. Multiple Paleopolyploidizations during the Evolution of the Compositae Reveal Parallel Patterns of Duplicate Gene Retention after Millions of Years. *Molecular Biology And Evolution*, 25(11), pp.2445–2455. Available at: http://mbe.oxfordjournals.org/cgi/doi/10.1093/molbev/msn187 [Accessed March 16, 2015].

Biémont, C. & Vieira, C., 2006. Genetics: junk DNA as an evolutionary force. *Nature*, 443(7111), pp.521–524.

Blanc, G., Barakat, A, Guyot, R., Cooke, R. & Delseny, M., 2000. Extensive duplication and reshuffling in the *Arabidopsis* genome. *The Plant Cell*, 12(7), pp.1093–1101.

Blanc, G. & Wolfe, K.H., 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *The Plant Cell*, 16(7), pp.1667–1678.

Blüthgen, N., Brand, K., Cajavec, B., Swat, M., Herzel, H. & Beule, D., 2005. Biological profiling of gene groups utilizing Gene Ontology. *Genome Informatics. International Conference On Genome Informatics*.

Bolger, A.M., Lohse, M., Usadel, B., Planck, M., Plant, M. & Mühlenberg, A., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, pp.1–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24695404.

Braasch, I., Brunet, F., Volff, J.-N.J.J.-N. & Schartl, M., 2009. Pigmentation pathway

evolution after whole-genome duplication in fish. *Genome Biology And Evolution*, 1, pp.479–93. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2839281&tool=pmcentrez&rendertype=abstract [Accessed October 5, 2011].

Braasch, I. & Postlethwait, J.H., 2012. Polyploidy in Fish and the Teleost Genome Duplication. In P. S. Soltis & D. E. Soltis, eds. *Polyploidy and Genome Evolution*. Springer Berlin Heidelberg, pp. 341–383.

Brawand, D., Wagner, C.E., Li, Y.I., Malinsky, M., Keller, I., Di Palma, F., et al., 2014. The genomic substrate for adaptive radiation in African cichlid fish. *Nature*. Available at: http://www.nature.com/doifinder/10.1038/nature13726 [Accessed September 3, 2014].

Brunet, F.G., Crollius, H.R., Paris, M., Aury, J.-M.M., Gibert, P., Jaillon, O., Laudet, V. & Robinson-Rechavi, M., 2006. Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Molecular Biology And Evolution*, 23(9), pp.1808–16. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16809621 [Accessed April 10, 2012].

Brunet, T.D.P. & Doolittle, W.F., 2015. Multilevel selection theory and the evolutionary functions of transposable elements. *Genome Biology And Evolution*, p.evv152. Available at: http://gbe.oxfordjournals.org/lookup/doi/10.1093/gbe/evv152.

Byrne, K.P. & Wolfe, K.H., 2007. Consistent patterns of rate asymmetry and gene loss indicate widespread neofunctionalization of yeast genes after whole-genome duplication. *Genetics*, 175(3), pp.1341–1350.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T.L., 2009. BLAST+: architecture and applications. *BMC Bioinformatics*, 10, p.421. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2803857&tool=pmcentrez&rendertype=abstract [Accessed July 9, 2014].

Cañestro, C. & Albalat, R., 2012. Transposon diversity is higher in amphioxus than in vertebrates: functional and evolutionary inferences. *Briefings In Functional Genomics*, 11(2). Available at: http://www.ncbi.nlm.nih.gov/pubmed/22389043 [Accessed March 10, 2012].

Capy, P., Gasperi, G., Biémont, C. & Bazin, C., 2000. Stress and transposable elements: Co-evolution or useful parasites? *Heredity*, 85(2), pp.101–106.

Carvalho, A., Delgado, M., Barão, A., Frescatada, M., Ribeiro, E., Pikaard, C.S., Viegas, W. & Neves, N., 2010. Chromosome and DNA methylation dynamics during meiosis in the autotetraploid Arabidopsis arenosa. *Sexual Plant Reproduction*, 23(1), pp.29–37.

Casacuberta, E. & González, J., 2013. The impact of transposable elements in environmental adaptation. *Molecular Ecology*, 22(6), pp.1503–1517.

Catchen, J.M., Amores, A., Hohenlohe, P., Cresko, W. & Postlethwait, J.H., 2011. Stacks:

building and genotyping Loci de novo from short-read sequences. *G3 (Bethesda, Md.)*, 1(3), pp.171–82. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3276136&tool=pmcentrez&rendertype=abstract [Accessed March 1, 2013].

Chao, D., Dilkes, B., Luo, H., Douglas, A., Yakubova, E., Lahner, B. & Salt, D.E., 2013. Polyploids exhibit higher potassium uptake and salinity tolerance in Arabidopsis. *Science*, 341(August), pp.658–659.

Charles, M., Belcram, H., Just, J., Huneau, C., Viollet, A., Couloux, A., Segurens, B., Carter, M., Huteau, V., Coriton, O., Appels, R., Samain, S. & Chalhoub, B., 2008. Dynamics and differential proliferation of transposable elements during the evolution of the B and A genomes of wheat. *Genetics*, 180(2), pp.1071–1086.

Chénais, B., Caruso, A., Hiard, S. & Casse, N., 2012. The impact of transposable elements on eukaryotic genomes: From genome size increase to genetic adaptation to stressful environments. *Gene*, 509(1), pp.7–15. Available at: http://dx.doi.org/10.1016/j.gene.2012.07.042.

Chiu, C., Amemiya, C., Dewar, K., Kim, C.-B., Ruddle, F.H. & Wagner, G.P., 2002. Molecular evolution of the *HoxA* cluster in the three major gnathostome lineages. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 99(8), pp.5492–7. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=122797&tool=pmcentrez&rendertype=abstract.

Chiu, C.-H., Dewar, K., Wagner, G.P., Takahashi, K., Ruddle, F., Ledje, C., Bartsch, P., Scemama, J.-L., Stellwag, E., Fried, C., Prohaska, S.J., Stadler, P.F. & Amemiya, C.T., 2004. Bichir *HoxA* cluster sequence reveals surprising trends in ray-finned fish genomic evolution. *Genome Research*, 14(1), pp.11–7. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=314268&tool=pmcentrez&rendertype=abstract [Accessed June 10, 2011].

Christoffels, A., Koh, E.G.L., Chia, J., Brenner, S., Aparicio, S. & Venkatesh, B., 2004. Fugu Genome Analysis Provides Evidence for a Whole-Genome Duplication Early During the Evolution of Ray-Finned Fishes. *Molecular Biology And Evolution*, 21(6), pp.1146–1151.

Chung, H., Bogwitz, M.R., McCart, C., Andrianopoulos, A., Ffrench-Constant, R.H., Batterham, P. & Daborn, P.J., 2007. Cis-regulatory elements in the accord retrotransposon result in tissue-specific expression of the *Drosophila melanogaster* insecticide resistance gene Cyp6g1. *Genetics*, 175(3), pp.1071–1077.

Cifuentes, M., Grandont, L., Moore, G., Chèvre, A.M. & Jenczewski, E., 2010. Genetic regulation of meiosis in polyploid species: New insights into an old question. *New Phytologist*, 186(1), pp.29–36.

Comai, L., 2005. The advantages and disadvantages of being polyploid. *Nature Reviews.*

Genetics, 6(11), pp.836–46. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16304599 [Accessed October 26, 2012].

Le Comber, S.C., Ainouche, M.L., Kovarik, A. & Leitch, A.R., 2010. Making a functional diploid: From polysomic to disomic inheritance. *New Phytologist*, 186(1), pp.113–122.

Le Comber, S.C. & Smith, C., 2004. Polyploidy in fishes: patterns and processes. *Biological Journal Of The Linnean Society*, 82(4), pp.431–442. Available at: http://onlinelibrary.wiley.com/doi/10.1111/j.1095-8312.2004.00330.x/full [Accessed October 28, 2011].

Conant, G.C., Birchler, J. A & Pires, J.C., 2014. Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. *Current Opinion In Plant Biology*, 19(June), pp.91–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24907529 [Accessed December 15, 2014].

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M. & Robles, M., 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics (Oxford, England)*, 21(18), pp.3674–6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16081474 [Accessed January 21, 2014].

Consortium, I.H.G., 2001. Initial sequencing and analysis of the human genome. *Nature*, 420(February), pp.520–562.

Crow, K.D., Amemiya, C.T., Roth, J. & Wagner, G.P., 2009. Hypermutability of *HoxA13A* and functional divergence from its paralog are associated with the origin of a novel developmental feature in zebrafish and related taxa (Cypriniformes). *Evolution; International Journal Of Organic Evolution*, 63(6), pp.1574–92. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19222565 [Accessed May 1, 2012].

Crow, K.D., Smith, C.D., Cheng, J.F., Wagner, G.P. & Amemiya, C.T., 2012. An independent genome duplication inferred from Hox paralogs in the American paddlefish-a representative basal ray-finned fish and important comparative reference. *Genome Biology And Evolution*, 4(9), pp.937–953.

Crow, K.D., Stadler, P.F., Lynch, V.J., Amemiya, C. & Wagner, G.P., 2006. The "fish-specific" *Hox* cluster duplication is coincident with the origin of teleosts. *Molecular Biology And Evolution*, 23(1), pp.121–36. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16162861 [Accessed March 6, 2012].

Crow, K.D. & Wagner, P., 2005. Proceedings of the SMBE Tri-National Young Investigators ' Workshop 2005 What Is the Role of Genome Duplication in the Evolution of Complexity and Diversity ? *Molecular Biology*.

Cummings, S.M., McMullan, M., Joyce, D. A. & van Oosterhout, C., 2010. Solutions for PCR, cloning and sequencing errors in population genetic analysis. *Conservation Genetics*, 11(3), pp.1095–1097.

Darriba, D., Toboada, G.L., Doallo, R. & Posada, D., 2012. jModelTest2: more models, new

heuristics and parallel computing. *Nature Methods2*, 9.

Davey, J.W., Hohenlohe, P. A, Etter, P.D., Boone, J.Q., Catchen, J.M. & Blaxter, M.L., 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews. Genetics*, 12(7), pp.499–510. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21681211 [Accessed February 28, 2013].

Dehal, P. & Boore, J.L., 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biology*, 3(10), p.e314. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1197285&tool=pmcentrez&rendertype=abstract [Accessed March 18, 2012].

Dernburg, A.F. & Karpen, G.H., 2002. A chromosome RNAissance. *Cell*, 111(2), pp.159–162.

van Dijk, E.L., Auger, H., Jaszczyszyn, Y. & Thermes, C., 2014. Ten years of next-generation sequencing technology. *Trends In Genetics*, pp.1–9. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0168952514001127 [Accessed August 7, 2014].

Donoghue, P.C.J. & Purnell, M. A, 2005. Genome duplication, extinction and vertebrate evolution. *Trends In Ecology & Evolution*, 20(6), pp.312–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16701387 [Accessed June 19, 2011].

Doolittle, W.F. & Sapienza, C., 1980. Selfish genes, the phenotype paradigm and genome evolution. *Nature*, 284(5757), pp.601–603.

Doyle, J.J., Flagel, L.E., Paterson, A.H., Rapp, R.A., Soltis, D.E., Soltis, P.S. & Wendel, J.F., 2008. Evolutionary genetics of genome merger and doubling in plants. *Annual Review Of Genetics*, 42(1), pp.443–461. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18983261.

Duboule, D., 2007. The rise and fall of *Hox* gene clusters. *Development (Cambridge, England)*, 134(14), pp.2549–2560.

Dujon, B., Sherman, D., Fischer, G., Durrens, P., Casaregola, S., Souciet, J.-L., et al., 2004. Genome evolution in yeasts. *Nature*, 430(6995), pp.35–44.

Durand, D., 2003. Vertebrate evolution: doubling and shuffling with a full deck. *Trends In Genetics*, 19(1), pp.2–5. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0168952502000082.

Durand, D. & Hoberman, R., 2006. Diagnosing duplications--can it be done? *Trends In Genetics : TIG*, 22(3), pp.156–164.

Eaton, D. a. R., 2013. PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. Available at: http://biorxiv.org/lookup/doi/10.1101/001081 [Accessed January 17, 2014].

Eaton, D.A.. & Ree, R.., 2013. Inferring Phylogeny and Introgression using RADseq Data : An Example from Flowering Plants (Pedicularis : Orobanchaceae). *Systematic Biology*,

0(0), pp.1–18.

Edgar, R.C., 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5), pp.1792–1797.

Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics (Oxford, England)*, 26(19), pp.2460–1. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20709691 [Accessed September 20, 2013].

Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C. & Knight, R., 2011. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, 27(16), pp.2194–2200.

Eilam, T., Anikster, Y., Millet, E., Manisterski, J. & Feldman, M., 2010. Genome Size in Diploids, Allopolyploids, and Autopolyploids of Mediterranean Triticeae. *Journal Of Botany*, 2010, pp.1–12.

Ekblom, R. & Galindo, J., 2011. Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, 107(1), pp.1–15. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3186121&tool=pmcentrez&rendertype=abstract [Accessed March 1, 2013].

Eschmeyer, W.N., 2013. Catalog of Fishes, California Academy of Sciences, San Francisco. Available at: http://research.calacademy.org/ichthyology/catalog.

Etter, P.D., Bassham, S., Hohenlohe, P.A., Johnson, E. & Cresko, W.A., 2011. SNP Discovery and Genotyping for Evolutionary Genetics Using RAD Sequencing. In V. Orgogozo & M. V. Rockman, eds. *Molecular Methods for Evolutionary Genetics*. Methods in Molecular Biology. Totowa, NJ: Humana Press. Available at: http://www.springerlink.com/index/10.1007/978-1-61779-228-1 [Accessed January 29, 2013].

Etter, P.D., Preston, J.L., Bassham, S., Cresko, W. A. & Johnson, E. a, 2011. Local de novo assembly of RAD paired-end contigs using short sequencing reads. *PloS One*, 6(4), p.e18561. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3076424&tool=pmcentrez&rendertype=abstract [Accessed January 29, 2013].

Ewing, B. & Green, P., 1998. Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. *Genome Research*, 8, pp.186–194.

Ewing, B., Hillier, L., Wendl, M.C. & Green, P., 1998. Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment. *Genome Research*, 8, pp.175–185.

Fawcett, J.A., Maere, S. & Van de Peer, Y., 2009. Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 106(14), pp.5737–42. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2667025&tool=pmcentrez&rendertype=abstract.

Ferraris, C.J., 2007. Checklist of catfishes, recent and fossil (Osteichthyes: Siluriformes), and catalogue of siluriform primary types. *Zootaxa*, 1418.

Feschotte, C., Jiang, N. & Wessler, S.R., 2002. Plant transposable elements: where genetics meets genomics. *Nature Reviews. Genetics*, 3(5), pp.329–41. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11988759 [Accessed February 24, 2014].

Feschotte, C. & Pritham, E.J., 2007. DNA transposons and the evolution of eukaryotic genomes. *Annual Review Of Genetics*, 41, pp.331–368.

Finnegan, D.J., 1989. Eukaryotic Transposable Elements and Genome Evolution. *Tig*, (4).

Finnegan, D.J., 1992. Transposable elements. *Current Opinion In Genetics & Development*, 2(6), pp.861–867.

Flagel, L.E. & Wendel, J.F., 2010. Evolutionary rate variation, genomic dominance and duplicate gene expression evolution during allotetraploid cotton speciation. *New Phytologist*, 186(1), pp.184–193.

Fontdevila, A., 2011. The Genome is Mobile. In *The Dynamic Genome - A Darwinian Approach*. Oxford University Press, pp. 80–115.

Freeling, M., 2009. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annual Review Of Plant Biology*, 60, pp.433–453.

Freeling, M., Woodhouse, M.R., Subramaniam, S., Turco, G., Lisch, D. & Schnable, J.C., 2012. Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. *Current Opinion In Plant Biology*, 15(2), pp.131–139. Available at: http://dx.doi.org/10.1016/j.pbi.2012.01.015.

Froese, R. & Pauly, D., 2015. Fishbase. Available at: www.fishbase.org.

Fultz, D., Choudury, S.G. & Slotkin, R.K., 2015. Silencing of active transposable elements in plants. *Current Opinion In Plant Biology*, 27, pp.67–76. Available at: http://dx.doi.org/10.1016/j.pbi.2015.05.027.

Furlong, R.F. & Holland, P.W.H., 2002. Were vertebrates octoploid? *Philosophical Transactions Of The Royal Society Of London. Series B, Biological Sciences*, 357(1420), pp.531–44. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1692965&tool=pmcentrez&rendertype=abstract [Accessed March 22, 2012].

Garcia de la Serrana, D., Mareco, E. A & Johnston, I. A, 2014. Systematic variation in the pattern of gene paralog retention between the teleost superorders Ostariophysi and Acanthopterygii. *Genome Biology And Evolution*, 6(4), pp.981–7. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4007551&tool=pmcentrez&rendertype=abstract [Accessed January 5, 2015].

Garcia-Fernàndez, J., Holland, P.W.H. & others, 1994. Archetypal organization of the

amphioxus *Hox* gene cluster. *Nature*, 370(6490), pp.563–566. Available at: http://teosinte.wisc.edu/gen677_pdfs/Garcia_Fernandez.pdf [Accessed March 29, 2012].

Garrison, E. & Marth, G., 2012. Haplotype-based variant detection from short-read sequencing. *ArXiv Preprint*, pp.1–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24136966.

Garsmeur, O., Schnable, J.C., Almeida, A., Jourda, C., D'Hont, A. & Freeling, M., 2014. Two evolutionarily distinct classes of paleopolyploidy. *Molecular Biology And Evolution*, 31(2), pp.448–454.

Gladman, S. & Seeman, T., 2012. VelvetOptimiser. Available at: http://bioinformatics.net.au/software.velvetoptimiser.shtml.

Glasauer, S.M.K. & Neuhauss, S.C.F., 2014. Whole-genome duplication in teleost fishes and its evolutionary consequences. *Molecular Genetics And Genomics : MGG*. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25092473 [Accessed August 9, 2014].

González, J., Karasov, T.L., Messer, P.W. & Petrov, D.A., 2010. Genome-wide patterns of adaptation to temperate environments associated with transposable elements in Drosophila. *PLoS Genetics*, 6(4), pp.33–35.

Gout, J.-F. & Lynch, M., 2015. Maintenance and loss of duplicated genes by dosage subfunctionalization. *Molecular Biology And Evolution*, 32(8), pp.2141–2148.

Grandbastien, M.A., 1998. Activation of plant retrotransposons under stress conditions. *Trends In Plant Science*, 3(5), pp.181–187.

Grandbastien, M.A., Audeon, C., Bonnivard, E., Casacuberta, J.M., Chalhoub, B., Costa, A.P.P., Le, Q.H., Melayah, D., Petit, M., Poncet, C., Tam, S.M., Van Sluys, M.A. & Mhiri, C., 2005. Stress activation and genomic impact of Tnt1 retrotransposons in Solanaceae. *Cytogenetic And Genome Research*, 110(1-4), pp.229–241.

Gregory, T.R., 2001. Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. *Biological Reviews Of The Cambridge Philosophical Society*, 76(1), pp.65–101.

Gregory, T.R., 2005a. Genome Size Evolution in Animals. In T. R. Gregory, ed. *The Evolution of the Genome*. Elsevier Academic Press, pp. 4–71.

Gregory, T.R., 2005b. The C-value enigma in plants and animals: A review of parallels and an appeal for partnership. *Annals Of Botany*, 95(1), pp.133–146.

Gregory, T.R. & Mable, B.K., 2005. Polyploidy in Animals. In T. R. Gregory, ed. *The Evolution of the Genome*. Elsevier Academic Press, pp. 428–517.

Gu, X., 2013. DIVERGE MANUAL version 3 . 0 (DetectIng Variability in Evolutionary Rates among GEnes).

Gu, X. & Vander Velden, K., 2002. DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family. *Bioinformatics (Oxford, England)*, 18(3),

pp.500–501.

Gu, X., Zou, Y., Su, Z., Huang, W., Zhou, Z., Arendsee, Z. & Zeng, Y., 2013. An update of DIVERGE software for functional divergence analysis of protein family. *Molecular Biology And Evolution*, 30(7), pp.1713–1719.

Guindon, S. & Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52(5), pp.696–704.

Guo, B., Gan, X. & He, S., 2010. *Hox* genes of the Japanese eel Anguilla japonica and Hox cluster evolution in teleosts. *Journal Of Experimental Zoology Part B: Molecular And Developmental Evolution*, 314 B(2), pp.135–147.

Hakes, L., Pinney, J.W., Lovell, S.C., Oliver, S.G. & Robertson, D.L., 2007. All duplicates are not equal: the difference between small-scale and genome duplication. *Genome Biology*, 8(10), p.R209.

Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., White, R., et al., 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, 32(Database issue), pp.D258–D261.

Hawkins, J.S., Kim, H., Nason, J.D., Wing, R. a & Wendel, J.F., 2006. Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium. *Genome Research*, (515), pp.1252–1261.

Henkel, C. V., Burgerhout, E., de Wijze, D.L., Dirks, R.P., Minegishi, Y., Jansen, H.J., Spaink, H.P., Dufour, S., Weltzien, F.A., Tsukamoto, K. & van den Thillart, G.E.E.J.M., 2012. Primitive duplicate *hox* clusters in the european eel's genome. *PLoS ONE*, 7(2).

Hoegg, S., Brinkmann, H., Taylor, J.S. & Meyer, A., 2004. Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. *Journal Of Molecular Evolution*, 59(2), pp.190–203. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15486693 [Accessed April 3, 2012].

Hohenlohe, P. A, Bassham, S., Etter, P.D., Stiffler, N., Johnson, E. a & Cresko, W. A, 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, 6(2), p.e1000862. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2829049&tool=pmcentrez&rendertype=abstract [Accessed January 28, 2013].

Hohenlohe, P.A., Catchen, J. & Cresko, W.A., 2012. Population Genomic Analysis of Model and Nonmodel Organisms Using Sequenced RAD Tags. In F. Pompanon & A. Bonin, eds. *Data Production and Analysis in Population Genomics SE - 14*. Methods in Molecular Biology. Humana Press, pp. 235–260. Available at: http://dx.doi.org/10.1007/978-1-61779-870-2_14.

Hu, G., Hawkins, J.S., Grover, C.E. & Wendel, J.F., 2010. The history and disposition of

transposable elements in polyploid Gossypium. *Genome / National Research Council Canada = Genome / Conseil National De Recherches Canada*, 53(8), pp.599–607.

Hughes, A.L., Silva, J. & Friedman, R., 2001. Ancient Genome Duplications Did Not Structure the Human Hox -Bearing Chromosomes Ancient Genome Duplications Did Not Structure the Human Hox -Bearing Chromosomes. *Genome Research*, 11, pp.771–780.

Hurst, G.D.D. & Werren, J.H., 2001. The role of selfish genetic elements in eukaryotic evolution. *Nat Rev Genet*, 2(8), pp.597–606. Available at: http://dx.doi.org/10.1038/35084545.

Innan, H. & Kondrashov, F., 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews. Genetics*, 11(2), pp.97–108.

Jaillon, O., Aury, J., Brunet, F. & Petit, J., 2004. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature*, 431(October). Available at: http://www.nature.com/nature/journal/v431/n7011/abs/nature03025.html [Accessed June 22, 2012].

Jardim, S.S., Schuch, A.P., Pereira, C.M. & Loreto, E.L.S., 2015. Effects of heat and UV radiation on the mobilization of transposon mariner-Mos1. *Cell Stress And Chaperones*. Available at: http://link.springer.com/10.1007/s12192-015-0611-2.

Jiang, Y., Lu, J., Peatman, E., Kucuktas, H., Liu, S., Wang, S., Sun, F. & Liu, Z., 2011. A pilot study for channel catfish whole genome sequencing and de novo assembly. *BMC Genomics*, 12(1), p.629. Available at: http://www.biomedcentral.com/1471-2164/12/629.

Jiao, Y., Wickett, N.J., Ayyampalayam, S., Chanderbali, A.S., Landherr, L., Ralph, P.E., Tomsho, L.P., Hu, Y., Liang, H., Soltis, P.S., Soltis, D.E., Clifton, S.W., Schlarbaum, S.E., Schuster, S.C., Ma, H., Leebens-Mack, J. & dePamphilis, C.W., 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature*, 473(7345), pp.97–100. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21478875 [Accessed August 6, 2014].

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O. & Walichiewicz, J., 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic And Genome Research*, 110(1-4), pp.462–467.

Kanazawa, A., Liu, B., Kong, F., Arase, S. & Abe, J., 2009. Adaptive evolution involving gene duplication and insertion of a novel Ty1/copia-like retrotransposon in soybean. *Journal Of Molecular Evolution*, 69(2), pp.164–175.

Kasahara, M., Naruse, K., Sasaki, S., Nakatani, Y., Qu, W., Ahsan, B., Yamada, T., Nagayasu, Y., Doi, K., Kasai, Y., Jindo, T., Kobayashi, D., Shimada, A., Toyoda, A., Kuroki, Y., Fujiyama, A., Sasaki, T., Shimizu, A., Asakawa, S., Shimizu, N., Hashimoto, S.-I., Yang, J., Lee, Y., Matsushima, K., Sugano, S., Sakaizumi, M., Narita, T., Ohishi, K., Haga, S., Ohta, F., Nomoto, H., Nogata, K., Morishita, T., Endo, T., Shin-I, T.,

Takeda, H., Morishita, S. & Kohara, Y., 2007. The medaka draft genome and insights into vertebrate genome evolution. *Nature*, 447(7145), pp.714–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17554307 [Accessed March 2, 2012].

Kawakami, K., Shima, A. & Kawakami, N., 2000. Identification of a functional transposase of the Tol2 element, an Ac-like element from the Japanese medaka fish, and its transposition in the zebrafish germ lineage. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 97(21), pp.11403–11408.

Kazazian, H.H., 2004. Mobile elements: drivers of genome evolution. *Science (New York, N.Y.)*, 303(5664), pp.1626–1632.

Keynes, R. & Krumlauf, R., 1994. *Hox* genes and regionalization of the nervous system. *Annual Review Of Neuroscience*, 17, pp.109–132. Available at: http://www.annualreviews.org/doi/pdf/10.1146/annurev.ne.17.030194.000545 [Accessed March 29, 2012].

Kidwell, M.G., 2005. Transposable Elements. In T. R. Gregory, ed. *The Evolution of the Genome*. pp. 165–221.

Kidwell, M.G., 2002. Transposable elements and the evolution of genome size in eukaryotes. *Genetica*, 115, pp.49–63.

Kidwell, M.G. & Lisch, D., 1997. Transposable elements as sources of variation in animals and plants. *Proc. Natl. Acad. Sci. USA*, 94(15), pp.7704–7711. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=33680&tool=pmcentrez&rendertype=abstract.

Kidwell, M.G. & Lisch, D.R., 2000. Transposable elements and host genome evolution. *Trends In Ecology And Evolution*, 15(3), pp.95–99.

Kim, M.Y. & Zilberman, D., 2014. DNA methylation as a system of plant genomic immunity. *Trends In Plant Science*, 19(5), pp.320–326. Available at: http://dx.doi.org/10.1016/j.tplants.2014.01.014.

Kim, Y.B., Oh, J.H., McIver, L.J., Rashkovetsky, E., Michalak, K., Garner, H.R., Kang, L., Nevo, E., Korol, A.B. & Michalak, P., 2014. Divergence of *Drosophila melanogaster* repeatomes in response to a sharp microclimate contrast in Evolution Canyon, Israel. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 111(29), pp.10630–5. Available at: http://www.pnas.org/cgi/doi/10.1073/pnas.1410372111\nhttp://www.ncbi.nlm.nih.gov/pubmed/25006263.

Kondrashov, F.A., Rogozin, I.B., Wolf, Y.I. & Koonin, E. V, 2002. Selection in the evolution of gene duplications. *Genome Biology*, 3(2). Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=65685&tool=pmcentrez&rendertype=abstract.

Krasnov, A., Koskinen, H., Afanasyev, S. & Mölsä, H., 2005. Transcribed Tc1-like

transposons in salmonid fish. *BMC Genomics*, 6, p.107.

Kuraku, S. & Meyer, A., 2009. The evolution and maintenance of *Hox* gene clusters in vertebrates and the teleost-specific genome duplication. *The International Journal Of Developmental Biology*, 53(5-6), pp.765–73. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19557682 [Accessed July 15, 2011].

Lee, S.-I. & Kim, N.-S., 2014. Transposable elements and genome size variations in plants. *Genomics & Informatics*, 12(3), pp.87–97. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25317107.

Lee, Y.C.G., 2015. The Role of piRNA-Mediated Epigenetic Silencing in the Population Dynamics of Transposable Elements in *Drosophila melanogaster*. *PLOS Genetics*, 11(6), p.e1005269. Available at: http://dx.plos.org/10.1371/journal.pgen.1005269.

Leggatt, R.A. & Iwama, G.K., 2004. Occurrence of polyploidy in the fishes. *Reviews In Fish Biology And Fisheries*, pp.237–246.

Leitch, I.J. & Bennett, M.D., 2004. Genome downsize in polyploids plants. *Biological Journal Of The Linnean Society*.

Levin, D.A., 1983. Polyploidy and novelty in flowering plants. *The American Naturalist*, 122(1), pp.1–25. Available at: http://www.jstor.org/stable/2461003.

Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Available at: http://arxiv.org/abs/1303.3997 [Accessed July 10, 2014].

Lin, R., Ding, L., Casola, C., Ripoll, D.R., Feschotte, C. & Wang, H., 2007. Transposase-Derived Transcription Factors Regulate Light Signaling in *Arabidopsis*. *Science*, 318(5854), pp.1302–1305.

Liu, B., Kanazawa, A., Matsumura, H., Takahashi, R., Harada, K. & Abe, J., 2008. Genetic redundancy in soybean photoresponses associated with duplication of the phytochrome A gene. *Genetics*, 180(2), pp.995–1007.

Ludwig, A., Belfiore, N.M., Pitra, C., Svirsky, V. & Jenneckens, I., 2001. Genome duplication events and functional reduction of ploidy levels in sturgeon (Acipenser, Huso and Scaphirhynchus). *Genetics*, 158(3), pp.1203–15. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1461728&tool=pmcentrez&rendertype=abstract.

Lynch, M. & Conery, J.S., 2000. The Evolutionary Fate and Consequences of Duplicate Genes. *Science*, 290(5494), pp.1151–1155. Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.290.5494.1151 [Accessed March 2, 2012].

Lynch, M. & Conery, J.S., 2003. The origins of genome complexity. *Science (New York, N.Y.)*, 302(5649), pp.1401–1404.

Ma, L.J., Ibrahim, A.S., Skory, C., Grabherr, M.G., Burger, G., Butler, M., Elias, M., Idnurm,

A., Lang, B.F., Sone, T., Abe, A., Calvo, S.E., Corrochano, L.M., Engels, R., Fu, J., Hansberg, W., Kim, J.M., Kodira, C.D., Koehrsen, M.J., Liu, B., Miranda-Saavedra, D., O'Leary, S., Ortiz-Castellanos, L., Poulter, R., Rodriguez-Romero, J., Ruiz-Herrera, J., Shen, Y.Q., Zeng, Q., Galagan, J., Birren, B.W., Cuomo, C. a. & Wickes, B.L., 2009. Genomic analysis of the basal lineage fungus *Rhizopus oryzae* reveals a whole-genome duplication. *PLoS Genetics*, 5(7).

Mable, B., Alexandrou, M. & Taylor, M., 2011. Genome duplication in amphibians and fish: an extended synthesis. *Journal Of Zoology*, 284, pp.151–182. Available at: http://books.google.com/books?hl=en&lr=&id=BvIKAAAAIAAJ&oi=fnd&pg=PA1&dq=Journal+of+Zoology&ots=_Z_IMxbWJf&sig=Vxd2dkv5tJruh-Npo1MFCP5yw1Q [Accessed October 5, 2011].

Mable, B.K., 2003. Breaking down taxonomic barriers in polyploidy research. *Trends In Plant Science*, 8(12), pp.582–90. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14659707 [Accessed February 4, 2015].

Mable, B.K., 2004. "Why Polyploidy is Rarer in Animals Than in Plants": myths and mechanisms. *Biological Journal Of The Linnean Society*, 136(6), pp.453–466.

Madden, T., 2003. The BLAST Sequence Analysis Tool. In *The NCBI Handbook*. Available at: http://www.ncbi.nlm.nih.gov/books/NBK21097/.

Maddison, D.. & Maddison, W.., 2014. Chromaseq: a Mesquite package for analyzing sequence chromatograms. Available at: http://mesquiteproject.org/packages/chromaseq/.

Marcet-Houben, M. & Gabaldón, T., 2015. Beyond the Whole-Genome Duplication: Phylogenetic Evidence for an Ancient Interspecies Hybridization in the Baker's Yeast Lineage. *PLOS Biology*, 13(8), p.e1002220. Available at: http://dx.plos.org/10.1371/journal.pbio.1002220.

Mardis, E.R., 2011. A decade's perspective on DNA sequencing technology. *Nature*, 470(7333), pp.198–203. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21307932 [Accessed January 28, 2013].

Mardis, E.R., 2008. The impact of next-generation sequencing technology on genetics. *Trends In Genetics : TIG*, 24(3), pp.133–41. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18262675 [Accessed January 29, 2013].

Martin, A., 2001. Letter To The Editor: Is Tetralogy True? Lack of Support for the "One-to-Four Rule." *Molecular Biology And Evolution*, 18(1), pp.89–93.

Mastretta-Yanes, A., Zamudio, S., Jorgensen, T.H., Arrigo, N., Alvarez, N., Pinero, D. & Emerson, B.C., 2014. Gene duplication, population genomics and species-level differentiation within a tropical mountain shrub. *Genome Biology And Evolution*, 3492225684(ext 283). Available at: http://gbe.oxfordjournals.org/cgi/doi/10.1093/gbe/evu205 [Accessed September 16, 2014].

Mayrose, I., Barker, M.S. & Otto, S.P., 2010. Probabilistic models of chromosome number evolution and the inference of polyploidy. *Systematic Biology*, 59(2), pp.132–44. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20525626 [Accessed August 18, 2011].

Mayrose, I., Zhan, S.H., Rothfels, C.J., Magnuson-Ford, K., Barker, M.S., Rieseberg, L.H. & Otto, S.P., 2011. Recently formed polyploid plants diversify at lower rates. *Science (New York, N.Y.)*, 333(6047), p.1257.

McCarthy, F.M., Wang, N., Magee, G.B., Nanduri, B., Lawrence, M.L., Camon, E.B., Barrell, D.G., Hill, D.P., Dolan, M.E., Williams, W.P., Luthe, D.S., Bridges, S.M. & Burgess, S.C., 2006. AgBase: a functional genomics resource for agriculture. *BMC Genomics*, 7, p.229.

McClintock, B., 1951. Chromosome organization and genic expression. *Cold Spring Harbor Symposia On Quantitative Biology*, 16, pp.13–47.

McClintock, B., 1984. The significance of responses of the genome to challenge. *Science*, 226, pp.792–801.

McDonald, J.F., Matyunina, Lilya, V., Wilson, S., Jordan, I.K., Bowen, N.J. & Miller, W.J., 1997. LTR retrotransposons and the evolution of eukaryotic enhancers. *Genetica*, 100, pp.3–13.

Mckenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Krenytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M.A., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. & DePristo, M.A., 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20, pp.1297–1303.

Metzker, M.L., 2010. Sequencing technologies [mdash] the next generation. *Nat Rev Genet*, 11(1), pp.31–46. Available at: http://dx.doi.org/10.1038/nrg2626.

Meyer, A. & Van de Peer, Y., 2005. From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). *BioEssays : News And Reviews In Molecular, Cellular And Developmental Biology*, 27(9), pp.937–45. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16108068 [Accessed March 27, 2012].

Mieczkowski, P.A., Lemoine, F.J. & Petes, T.D., 2006. Recombination between retrotransposons as a source of chromosome rearrangements in the yeast Saccharomyces cerevisiae. *DNA Repair*, 5(9-10), pp.1010–1020.

Milne, I., Stephen, G., Bayer, M., Cock, P.J. a, Pritchard, L., Cardle, L., Shaw, P.D., Marshall, D., Shawand, P.D. & Marshall, D., 2013. Using tablet for visual exploration of second-generation sequencing data. *Briefings In Bioinformatics*, 14(2), pp.193–202. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22445902 [Accessed March 24, 2015].

Mirsky, A.E. & Ris, H., 1951. The desoxyribonucleic acid content of animal cells and its

evolutionary significance. *The Journal Of General Physiology*, 34(4), pp.451–462.

Moghe, G.D. & Shiu, S.-H., 2014. The causes and molecular consequences of polyploidy in flowering plants. *Annals Of The New York Academy Of Sciences*, 1320(1), pp.16–34. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24903334 [Accessed January 6, 2015].

Montgomery, E.A., Huang, S.M., Langley, C.H. & Judd, B.H., 1991. Chromosome rearrangement by ectopic recombination in *Drosophila melanogaster*: Genome structure and evolution. *Genetics*, 129(4), pp.1085–1098.

Morel, G., Sterck, L., Swennen, D., Marcet-Houben, M., Onesime, D., Levasseur, A., Jacques, N., Mallet, S., Couloux, A., Labadie, K., Amselem, J., Beckerich, J.-M., Henrissat, B., Van de Peer, Y., Wincker, P., Souciet, J.-L., Gabaldón, T., Tinsley, C.R. & Casaregola, S., 2015. Differential gene retention as an evolutionary mechanism to generate biodiversity and adaptation in yeasts. *Scientific Reports*, 5(January), p.11571. Available at: http://www.nature.com/doifinder/10.1038/srep11571.

Mungpakdee, S., Seo, H.-C., Angotzi, A.R., Dong, X., Akalin, A. & Chourrout, D., 2008. Differential evolution of the 13 *Atlantic salmon Hox* clusters. *Molecular Biology And Evolution*, 25(7), pp.1333–43. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18424774 [Accessed August 18, 2011].

Nandi, S., Peatman, E., Xu, P., Wang, S., Li, P. & Liu, Z., 2007. Repeat structure of the catfish genome: A genomic and transcriptomic assessment of Tc1-like transposon elements in channel catfish (*Ictalurus punctatus*). *Genetica*, 131(1), pp.81–90.

Nekrutenko, A. & Li, W.H., 2001. Transposable elements are found in a large number of human protein-coding genes. *Trends In Genetics*, 17(11), pp.619–621.

Nelson, J.S., 1994. *Fishes of the World* 3rd ed., New York: John Wiley, Sons, Inc.

O'Neil, S.T. & Emrich, S.J., 2012. Haplotype and minimum-chimerism consensus determination using short sequence data. *BMC Genomics*, 13 Suppl 2(Suppl 2), p.S4. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3394418&tool=pmcentrez&rendertype=abstract [Accessed December 4, 2014].

O'Neil, S.T. & Emrich, S.J., 2011. Robust haplotype reconstruction of eukaryotic read data with Hapler. *2011 IEEE 1st International Conference On Computational Advances In Bio And Medical Sciences (ICCABS)*, pp.141–146. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5729869.

Ogden, R., Gharbi, K., Mugue, N., Martinsohn, J., Senn, H., Davey, J.W., Pourkazemi, M., McEwing, R., Eland, C., Vidotto, M., Sergeev, A & Congiu, L., 2013. Sturgeon conservation genomics: SNP discovery and validation using RAD sequencing. *Molecular Ecology*, 22(11), pp.3112–23. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23473098 [Accessed September 20, 2013].

Ohno, S., Ohno, M. & Ohno, S., 1970. *Evolution by Gene Duplication*, Springer-Verlag.

Oliveira, C., Almeida-Toledo, L.F., Mori, L. & Toledo-Filho, S.A., 1993. Cytogenetic and DNA content studies of armoured catfishes of the genus *Corydoras* (Pisces, Siluriformes, Callichtyhyidae). *Revista Brasil Genetica*, 16(3), pp.617–629.

Oliveira, C., Almeida-Toledo, L.F., Mori, L. & Toledo-Filho, S.A., 1992. Extensive chromosomal rearrangements and nuclear DNA content changes in the evolution of the armoured catfishes genus *Corydoras* (Pisces, Siluriformes, Callichthyidae). *Journal Of Fish Biology*, 40(3), pp.419–431. Available at: http://doi.wiley.com/10.1111/j.1095-8649.1992.tb02587.x.

Oliver, K.R. & Greene, W.K., 2009. Transposable elements: Powerful facilitators of evolution. *BioEssays*, 31(7), pp.703–714.

Oliver, K.R., McComb, J. a. & Greene, W.K., 2013. Transposable Elements: Powerful Contributors to Angiosperm Evolution and Diversity. *Genome Biology And Evolution*, 5(10), pp.1–35. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24065734 [Accessed September 27, 2013].

Orgel, L.E. & Crick, F.H.C., 1980. Selfish DNA: the ultimate parasite. *Nature*, 284, pp.604–607.

Oswald, B.P. & Nuismer, S.L., 2007. Neopolyploidy and pathogen resistance. *Proceedings. Biological Sciences / The Royal Society*, 274(1624), pp.2393–7. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2274975&tool=pmcentrez&rendertype=abstract [Accessed March 12, 2012].

Ota, R.R., Message, H.J., da Graça, W.J., Pavanelli, C.S., Júnio da Graça, W. & Pavanelli, C.S., 2015. Neotropical Siluriformes as a Model for Insights on Determining Biodiversity of Animal Groups. *Plos One*, 10(7), p.e0132913. Available at: http://dx.plos.org/10.1371/journal.pone.0132913.

Otto, S. & Whitton, J., 2000. Polyploid incidence and evolution. *Annual Review Of Genetics*, 34, pp.401–7. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.150.5197&rep=rep1&type=pdf [Accessed October 14, 2011].

Otto, S.P., 2007. The evolutionary consequences of polyploidy. *Cell*, 131(3), pp.452–62. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17981114 [Accessed March 6, 2012].

Pardue, M.-L. & DeBaryshe, P.G., 2011. Adapting to life at the end of the line: How Drosophila telomeric retrotransposons cope with their job. *Mobile Genetic Elements*, 1(2), pp.128–134.

Parisod, C., Holderegger, R. & Brochmann, C., 2010. Evolutionary consequences of autopolyploidy. *The New Phytologist*, 186(1), pp.5–17. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20070540 [Accessed March 11, 2012].

Parisod, C., Salmon, A., Zerjal, T., Tenaillon, M., Grandbastien, M.A. & Ainouche, M., 2009. Rapid structural and epigenetic reorganization near transposable elements in hybrid and allopolyploid genomes in Spartina. *New Phytologist*, 184(4), pp.1003–1015.

Paterson, A.H., Chapman, B.A., Kissinger, J.C., Bowers, J.E., Feltus, F.A. & Estill, J.C., 2006. Many gene and domain families have convergent fates following independent whole-genome duplication events in *Arabidopsis*, *Oryza*, *Saccharomyces* and *Tetraodon*. *Trends In Genetics : TIG*, 22(11), pp.597–602. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16979781.

Van de Peer, Y., Maere, S. & Meyer, A., 2009. The evolutionary significance of ancient genome duplications. *Nature Reviews. Genetics*, 10(10), pp.725–32. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19652647 [Accessed July 28, 2011].

Pegueroles, C., Laurie, S. & Albà, M.M., 2013. Accelerated evolution after gene duplication: A time-dependent process affecting just one copy. *Molecular Biology And Evolution*, 30(8), pp.1830–1842.

Pocwierz-Kotus, A., Burzynski, A. & Wenne, R., 2007. Family of Tc1-like elements from fish genomes and horizontal transfer. *Gene*, 390(1-2), pp.243–251.

Postlethwait, J., Amores, A., Cresko, W., Singer, A. & Yan, Y.-L., 2004. Subfunction partitioning, the teleost radiation and the annotation of the human genome. *Trends In Genetics : TIG*, 20(10), pp.481–90. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15363902 [Accessed March 6, 2012].

Prohaska, S.J. & Stadler, P.F., 2004. The duplication of the *Hox* gene clusters in teleost fishes. *Theory In Biosciences = Theorie In Den Biowissenschaften*, 123(1), pp.89–110. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18202881.

Putnam, N.H., Butts, T., Ferrier, D.E.K., Furlong, R.F., Hellsten, U., Kawashima, T., Robinson-Rechavi, M., Shoguchi, E., Terry, A., Yu, J., Benito-Gutiérrez, E.L., Dubchak, I., Garcia-Fernàndez, J., Gibson-Brown, J.J., Grigoriev, I. V, Horton, A.C., de Jong, P.J., Jurka, J., Kapitonov, V. V, Kohara, Y., Kuroki, Y., Lindquist, E., Lucas, S., Osoegawa, K., Pennacchio, L.A., Salamov, A.A., Satou, Y., Sauka-Spengler, T., Schmutz, J., Shin-I, T., Toyoda, A., Bronner-Fraser, M., Fujiyama, A., Holland, L.Z., Holland, P.W.H., Satoh, N. & Rokhsar, D.S., 2008. The amphioxus genome and the evolution of the chordate karyotype. *Nature*, 453(7198), pp.1064–1071.

Rabosky, D.L., Santini, F., Eastman, J., Smith, S.A., Sidlauskas, B., Chang, J. & Alfaro, M.E., 2013. Rates of speciation and morphological evolution are correlated across the largest vertebrate radiation. *Nature Communications*, 4, pp.1–8. Available at: http://dx.doi.org/10.1038/ncomms2958\npapers2://publication/doi/10.1038/ncomms2958.

Radice, A.D., Bugaj, B., Fitch, D.H.A. & Emmons, S.W., 1994. Widespread occurrence of the Tc1 transposon family: Tc1-like transposons from teleost fish. *MGG Molecular & General Genetics*, 244(6), pp.606–612.

Rambaut, A., 2015. FigTree. Available at: http://tree.bio.ed.ac.uk/software/figtree/ [Accessed February 4, 2015].

Recknagel, H., Elmer, K.R. & Meyer, A., 2013. A hybrid genetic linkage map of two ecologically and morphologically divergent Midas cichlid fishes (Amphilophus spp.) obtained by massively parallel DNA sequencing (ddRADSeq). *G3 (Bethesda, Md.)*, 3(1), pp.65–74. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3538344&tool=pmcentrez&rendertype=abstract [Accessed April 21, 2015].

Renny-Byfield, S. & Wendel, J.F., 2014. Doubling down on genomes: Polyploidy and crop plants. *American Journal Of Botany*, 101, pp.1–15. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25090999 [Accessed August 9, 2014].

Rensing, S.A., Ick, J., Fawcett, J.A., Lang, D., Zimmer, A., Van de Peer, Y. & Reski, R., 2007. An ancient genome duplication contributed to the abundance of metabolic genes in the moss Physcomitrella patens. *BMC Evolutionary Biology*, 7, p.130.

Robertson, H.M., 1995. The Tc1-mariner superfamily of transposons in animals. *Journal Of Insect Physiology*, 41(2), pp.99–105.

Ronfort, J., 1999. The mutation load under tetrasomic inheritance and its consequences for the evolution of the selfing rate in autotetraploid species. *Genetical Research*, 74(1), pp.31–42. Available at: http://www.journals.cambridge.org/abstract_S0016672399003845.

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. a. & Huelsenbeck, J.P., 2012. Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, 61(3), pp.539–542. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3329765&tool=pmcentrez&rendertype=abstract [Accessed July 10, 2014].

RStudio, 2012. RStudio: Integrated development environment for R. Available at: http://www.rstudio.org/.

Santini, F., Harmon, L.J., Carnevale, G. & Alfaro, M.E., 2009. Did genome duplication drive the origin of teleosts? A comparative study of diversification in ray-finned fishes. *BMC Evolutionary Biology*, 9, p.194. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2743667&tool=pmcentrez&rendertype=abstract [Accessed June 17, 2011].

Scannell, D.R., Byrne, K.P., Gordon, J.L., Wong, S. & Wolfe, K.H., 2006. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature*, 440(7082), pp.341–345.

Scannell, D.R., Frank, A.C., Conant, G.C., Byrne, K.P., Woolfit, M. & Wolfe, K.H., 2007. Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. *Proceedings Of The National Academy Of*

Sciences Of The United States Of America, 104(20), pp.8397–8402.

Scheel, J.., Simonsen, V. & Gyldenholm, A.O., 1972. The karyotypes and some electrophoretic patterns of fourteen species of the genus Corydoras. *Journal Of Zoological Systematics And Evolutionary Research*, 10, pp.144–152.

Schnable, P., Ware, D., Fulton, R. & Stein, J., 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science*, 326(5956), pp.1112–1115. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19965430\nhttp://www.sciencemag.org/content/326/5956/1112.short.

Schneider, G.F. & Dekker, C., 2012. DNA sequencing with nanopores. *Nature Biotechnology*, 30(4), pp.326–328. Available at: http://dx.doi.org/10.1038/nbt.2181.

Schrader, L., Kim, J.W., Ence, D., Zimin, A., Klein, A., Wyschetzki, K., Weichselgartner, T., Kemena, C., Stökl, J., Schultner, E., Wurm, Y., Smith, C.D., Yandell, M., Heinze, J., Gadau, J. & Oettler, J., 2014. Transposable element islands facilitate adaptation to novel environments in an invasive species. *Nature Communications*, 5, p.5495. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4284661&tool=pmcentrez&rendertype=abstract [Accessed February 26, 2015].

Sémon, M. & Wolfe, K.H., 2007. Reciprocal gene loss between *Tetraodon* and zebrafish after whole genome duplication in their ancestor. *Trends In Genetics*, 23(3), pp.108–112.

Shoemaker, R.C., Schlueter, J. & Doyle, J.J., 2006. Paleopolyploidy and gene duplication in soybean and other legumes. *Current Opinion In Plant Biology*, 9(2), pp.104–109.

Smit, A., Hubley, R. & Green, P., RepeatMasker Open-4.0. Available at: <http://www.repeatmasker.org>.

Soltis, D.E., Albert, V.A., Leebens-Mack, J., Bell, C.D., Paterson, A.H., Zheng, C., Sankoff, D., DePamphilis, C.W., Wall, P.K. & Soltis, P.S., 2009. Polyploidy and angiosperm diversification. *American Journal Of Botany*, 96(1), pp.336–348.

Soltis, D.E., Segovia-Salcedo, M.C., Jordon-Thaden, I., Majure, L., Miles, N.M., Mavrodiev, E. V., Mei, W., Cortez, M.B., Soltis, P.S. & Gitzendanner, M.A., 2014. Are polyploids really evolutionary dead-ends (again)? A critical reappraisal of Mayrose et al. (2011). *New Phytologist*, 202(4), pp.1105–1117.

Soltis, P.S. & Soltis, D.E., 2000. The role of genetic and genomic attributes in the success of polyploids. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 97(13), pp.7051–7057.

Stamatakis, A., 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9), pp.1312–1313.

Stapley, J., Reger, J., Feulner, P.G.D., Smadja, C., Galindo, J., Ekblom, R., Bennison, C., Ball, A.D., Beckerman, A.P. & Slate, J., 2010. Adaptation genomics: the next generation. *Trends In Ecology & Evolution*, 25(12), pp.705–12. Available at:

http://www.ncbi.nlm.nih.gov/pubmed/20952088 [Accessed February 10, 2013].

Stebbins, G.L., 1971. *Chromosome Evolution in higher plants*, Hodder & Stoughton Educ.

Stebbins, G.L., 1950. *Variation and Evolution in Plants*, Columbia University Press.

Stern, A., Doron-Faigenboim, A., Erez, E., Martz, E., Bacharach, E. & Pupko, T., 2007. Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach. *Nucleic Acids Research*, 35(Web Server issue), pp.W506–11. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1933148&tool=pmcentrez&rendertype=abstract [Accessed February 4, 2015].

Sun, C., Shepard, D.B., Chong, R.A., Arriaza, J.L., Hall, K., Castoe, T.A., Feschotte, C., Pollock, D.D. & Mueller, R.L., 2012. LTR retrotransposons contribute to genomic gigantism in plethodontid salamanders. *Genome Biology And Evolution*, 4(2), pp.168–183.

Sunnucks, P. & Hales, D.F., 1996. Numerous transposed sequences of mitochondrial cytochrome oxidase I-II in aphids of the genus Sitobion (Hemiptera: Aphididae). *Molecular Biology And Evolution*, 13(3), pp.510–524. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8742640.

Tatusov, T. & Tatusov, R., 2015. NCBI ORF Finder (Open Reading Frame Finder). Available at: http://www.ncbi.nlm.nih.gov/projects/gorf/.

Taylor, J.S., Braasch, I., Frickey, T., Meyer, A. & Van de Peer, Y., 2003. Genome duplication, a trait shared by 22000 species of ray-finned fish. *Genome Research*, 13(3), pp.382–90. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=430266&tool=pmcentrez&rendertype=abstract [Accessed March 28, 2012].

Taylor, J.S., Van de Peer, Y., Braasch, I. & Meyer, a, 2001. Comparative genomics provides evidence for an ancient genome duplication event in fish. *Philosophical Transactions Of The Royal Society Of London. Series B, Biological Sciences*, 356(1414), pp.1661–79. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1088543&tool=pmcentrez&rendertype=abstract [Accessed April 2, 2012].

Taylor, J.S., Peer, Y. Van De & Meyer, A., 2001. Genome duplication , divergent resolution and speciation. *Trends In Genetics*, 17(6), pp.299–301.

Taylor, J.S. & Raes, J., 2004. Duplication and divergence: the evolution of new genes and old ideas. *Annual Review Of Genetics*, 38, pp.615–643.

Taylor, J.S. & Raes, J., 2005. Small-Scale Gene Duplications. In T. R. Gregory, ed. *The Evolution of the Genome*. Elsevier Academic Press, pp. 289–328.

The Arabidopsis Genome Initiative, 2000. Analysis of the genome sequence of the flowering

plant Arabidopsis thaliana. *Nature*, 408(6814), pp.796–815. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11130711.

Thomas, C.A., 1971. The genetic organization of chromosomes. *Annual Review Of Genetics*, 5, pp.237–256.

Tsigenopoulos, C.S., Ráb, P., Naran, D. & Berrebi, P., 2002. Multiple origins of polyploidy in the phylogeny of southern African barbs (Cyprinidae) as inferred from mtDNA markers. *Heredity*, 88(6), pp.466–473.

Turner, B., Diffoot, N. & Rasch, E., 1992. The challichtyid catfish *Corydoras aeneus* is an unresolved diploid-tetraploid sibling species complex. *Ichthyological Exploration Of Freshwaters*, 3, pp.17–23.

Vamosi, J.C. & Dickinson, T.A., 2006. Polyploidy and Diversification: A Phylogenetic Investigation in Rosaceae. *International Journal Of Plant Sciences*, 167(2), pp.349–358. Available at: http://www.jstor.org/stable/10.1086/499251.

Vandepoele, K., De Vos, W., Taylor, J.S., Meyer, A. & Van de Peer, Y., 2004. Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 101(6), pp.1638–1643.

Volff, J.N., 2006. Turning junk into gold: Domestication of transposable elements and the creation of new genes in eukaryotes. *BioEssays*, 28(9), pp.913–922.

Volff, J.-N., 2005. Genome evolution and biodiversity in teleost fish. *Heredity*, 94(3), pp.280–94. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15674378 [Accessed March 6, 2012].

Wagner, G.P. & Lynch, V.J., 2010. *Hox* cluster duplications and the opportunity for evolutionary novelties. *Current Biology : CB*, 20(2), pp.R48–52. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20129035.

Wagner, G.P., Takahashi, K., Lynch, V., Prohaska, S.J., Fried, C., Stadler, P.F. & Amemiya, C., 2005. Molecular evolution of duplicated ray finned fish *HoxA* clusters: increased synonymous substitution rate and asymmetrical co-divergence of coding and non-coding sequences. *Journal Of Molecular Evolution*, 60(5), pp.665–76. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15983874 [Accessed August 15, 2012].

Wang, J., Tian, L., Lee, H.-S., Wei, N.E., Jiang, H., Watson, B., Madlung, A., Osborn, T.C., Doerge, R.W., Comai, L. & Chen, Z.J., 2006. Genomewide nonadditive gene regulation in *Arabidopsis* allotetraploids. *Genetics*, 172(1), pp.507–17. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1456178&tool=pmcentrez&rendertype=abstract [Accessed March 29, 2012].

Wang, Y., Wang, X. & Paterson, A.H., 2012. Genome and gene duplications and gene expression divergence: A view from plants. *Annals Of The New York Academy Of Sciences*, 1256(1), pp.1–14.

Warren, I. A, Ciborowski, K.L., Casadei, E., Hazlerigg, D.G., Martin, S., Jordan, W.C. & Sumner, S., 2014. Extensive local gene duplication and functional divergence among paralogs in *Atlantic salmon*. *Genome Biology And Evolution*, 6(7), pp.1790–805. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4122929&tool=pmcentrez&rendertype=abstract [Accessed February 6, 2015].

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P. & Schulman, A.H., 2007. A unified classification system for eukaryotic transposable elements. *Nature Reviews. Genetics*, 8(12), pp.973–982.

Wickham, H., 2009. *ggplot2: Elegant Graphics for Data Analysis*, Springer.

Wittbrodt, J., Meyer, A. & Schartl, M., 1998. More genes in fish? *BioEssays*, 20(6), pp.511–515. Available at: http://doi.wiley.com/10.1002/(SICI)1521-1878(199806)20:6<511::AID-BIES10>3.0.CO;2-3.

Wolfe, K., 2000. Robustness - it's not where you think it is. *Nature Genetics*, 25(may 2000), pp.3–4.

Wolfe, K.H., 2001. Yesterday's polyploids and the mystery of diploidization. *Nature Reviews. Genetics*, 2(5), pp.333–341.

Wood, T.E., Takebayashi, N., Barker, M.S., Mayrose, I., Greenspoon, P.B. & Rieseberg, L.H., 2009. The frequency of polyploid speciation in vascular plants. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 106(33), pp.13875–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2728988&tool=pmcentrez&rendertype=abstract.

Woodhouse, M.R., Schnable, J.C., Pedersen, B.S., Lyons, E., Lisch, D., Subramaniam, S. & Freeling, M., 2010. Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homologs. *PLoS Biology*, 8(6).

Yang, Z., Nielsen, R., Goldman, N. & Pedersen, A., 2000. Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites. *Genetics*, 155(1), pp.431–449.

Yant, L., Hollister, J.D., Wright, K.M., Arnold, B.J., Higgins, J.D., Franklin, F.C.H. & Bomblies, K., 2013. Meiotic Adaptation to Genome Duplication in *Arabidopsis arenosa*. *Current Biology*, 23(21), pp.2151–2156. Available at: http://dx.doi.org/10.1016/j.cub.2013.08.059.

Yoshida, K., Schuenemann, V.J., Cano, L.M., Pais, M., Mishra, B., Sharma, R., Lanz, C., Martin, F.N., Kamoun, S., Krause, J., Thines, M., Weigel, D. & Burbano, H. a, 2013. The rise and fall of the *Phytophthora infestans* lineage that triggered the Irish potato famine. *ELife*, 2, p.e00731. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3667578&tool=pmcentrez&r

endertype=abstract [Accessed May 28, 2014].

Yuan, J., He, Z., Yuan, X., Jiang, X., Sun, X. & Zou, S., 2010. Speciation of polyploid Cyprinidae fish of common carp, crucian carp, and silver crucian carp derived from duplicated *Hox* genes. *Journal Of Experimental Zoology. Part B, Molecular And Developmental Evolution*, 314(6), pp.445–56. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20700889 [Accessed August 8, 2011].

Zerbino, D.R., 2010. Using the Velvet de novo assembler for short-read sequencing technologies. *Current Protocols In Bioinformatics*, (SUPPL. 31), pp.1–13.

Zerbino, D.R. & Birney, E., 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5), pp.821–829. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2336801&tool=pmcentrez&rendertype=abstract [Accessed July 9, 2014].

Zhan, S.H., Glick, L., Tsigenopoulos, C.S., Otto, S.P. & Mayrose, I., 2014. Comparative analysis reveals that polyploidy does not decelerate diversification in fish. *Journal Of Evolutionary Biology*, 27(2), pp.391–403.

Zhou, L., Mitra, R., Atkinson, P.W., Hickman, A.B., Dyda, F. & Craig, N.L., 2004. Transposition of hAT elements links transposable elements and V(D)J recombination. *Nature*, 432(7020), pp.995–1001.

Zuccolo, A., Sebastian, A., Talag, J., Yu, Y., Kim, H., Collura, K., Kudrna, D. & Wing, R.A., 2007. Transposable element distribution, abundance and role in genome size variation in the genus Oryza. *BMC Evolutionary Biology*, 7, p.152.