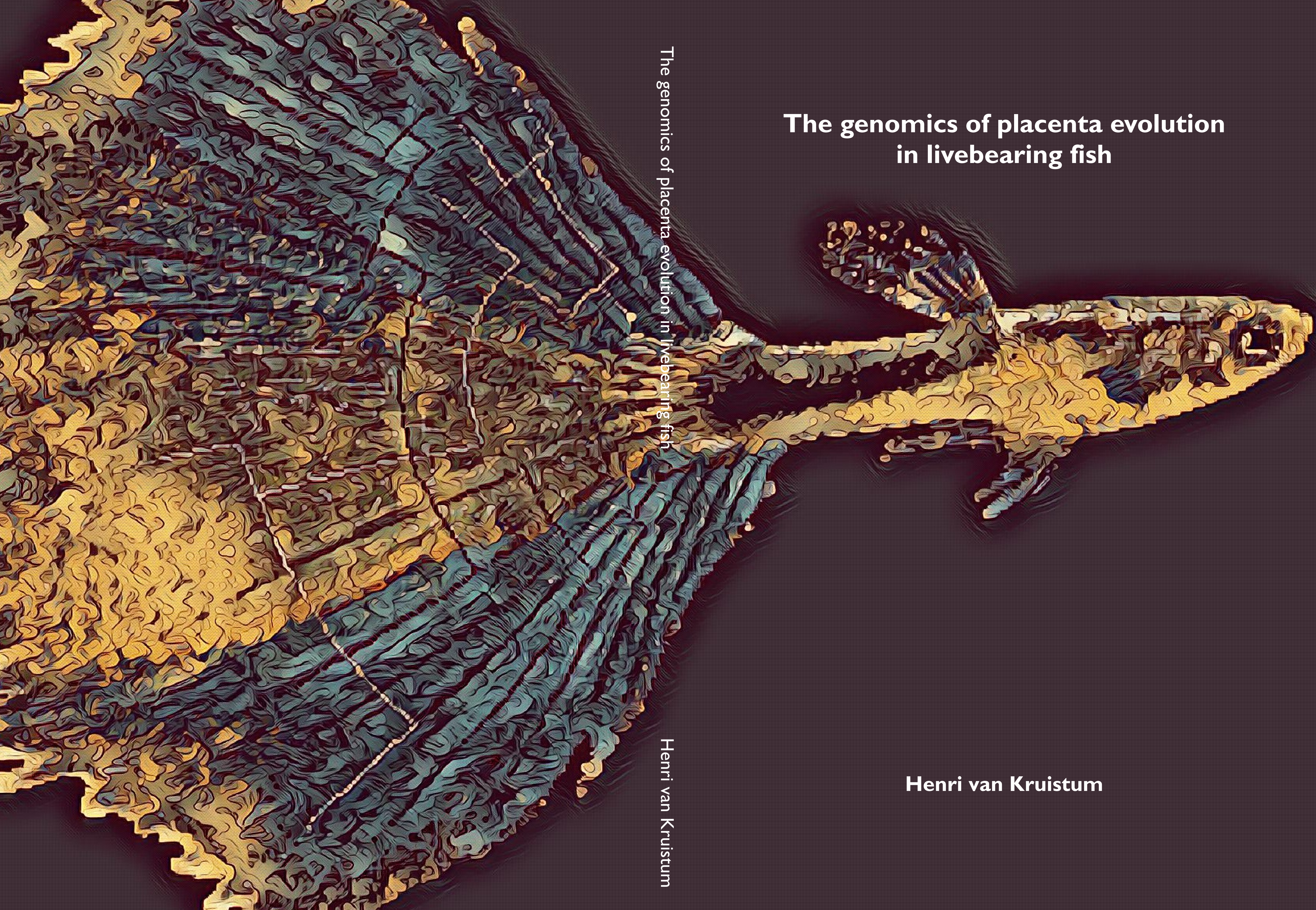

The genomics of placenta evolution in livebearing fish

Henri van Kruistum

The genomics of placenta evolution in livebearing fish
Henri van Kruistum

Propositions

1. The evolution of new vertebrate organs is driven by a change in the regulation of genes involved in early development.
(this thesis)

2. The computational cost of whole genome alignment is currently the limiting step for progress in the field of comparative genomics.
(this thesis)

3. Research money is better used to hire people that analyze existing omics data, than to generate more omics data.

4. A PhD project should only be allowed to be funded by private companies if all associated data is made publicly available.

5. An opinion being insulting does not warrant censorship.

6. Current social media platforms encourage polarization.

Propositions belonging to the thesis, entitled:

The genomics of placenta evolution in livebearing fish

Henri van Kruistum
Wageningen, 6 April 2022

The genomics of placenta evolution in livebearing fish

Henri van Kruistum

Thesis committee

Promotors
Prof. Dr M.A.M. Groenen
Professor of Animal Breeding and Genomics
Wageningen University & Research

Prof. Dr J.L. van Leeuwen
Professor of Experimental Zoology
Wageningen University & Research

Co-promotors
Dr H.-J.W.C. Megens
Assistant professor, Animal Breeding and Genomics
Wageningen University & Research

Dr B.J.A Pollux
Associate professor, Experimental Zoology
Wageningen University & Research

Other members
Prof. Dr D.J. Macqueen, University of Edinburgh, Scotland
Prof. Dr G.G. Rosenthal, Padova University, Italy
Prof. Dr D. de Ridder, Wageningen University & Research
Prof. Dr B. Wertheim, University of Groningen

This research was conducted under the auspices of the Graduate School of Wageningen Institute of Animal Sciences (WIAS).

The genomics of placenta evolution in livebearing fish

Henri van Kruistum

Thesis

submitted in fulfillment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Wednesday 6 April 2022
at 4. p.m. in the Aula.

Henri van Kruistum
The genomics of placenta evolution in livebearing fish

PhD thesis, Wageningen University, the Netherlands (2021)
With references, with summary in English

ISBN: 978-94-6447-037-6
DOI: https://doi.org/10.18174/558081

Abstract

During vertebrate evolution, complex organs have evolved several times. The development of these organs is encoded in the genome. However, it is currently unclear how new complex organs, consisting of multiple interlocking parts, can evolve as a result of genomic change. In this thesis, I aim to find the genomic basis of one such newly evolved organ: the placenta in the livebearing fish family Poeciliidae. In this family, a placenta has evolved nine times independently, allowing for the investigation of multiple evolutionary origins of the same organ within a single group of species. First, I sequence and assemble the genomes of both placental and non-placental poeciliid species. Then, I compare the genomes of placental species with the genomes of non-placental species, aiming to find consistent genomic differences between placental and non-placental species that can be associated with placenta evolution. I show that indeed, placental species show consistent mutations in both protein-coding and regulatory regions of the genome. Protein-coding mutations occur mainly around structural and metabolic genes, while regulatory changes occur mainly around developmental genes. I also show that, contrary to some predictions, gene duplications are not associated with placenta evolution in poeciliid fish. Finally, I show that allele-specific DNA methylation is present in the poeciliid fish *Poeciliopsis gracilis*, and that its inheritance is non-random but instead depends on parent-of-origin of the methylated allele, suggesting genomic imprinting. Together I provide a comprehensive overview of genome evolution in the fish family Poeciliidae, and provide new insights into the evolution of complex traits.

Contents

1

General introduction

1.1 DNA and the evolution of life

For centuries, humans have been amazed by the complexity and diversity of life on earth, and the question of where it comes from has likely been on millions of minds before us. For scientists, the most famous early attempt of answering this question is that of Charles Darwin. In his work on the origin of species, he proposed the stepwise adaptation of species to their environment through natural selection, the principle that now is the foundation of the field of evolutionary biology (Darwin 1859). Although Darwin accurately described the process of adaptation of species to their environment, the mechanism by which these adaptations were stored and inherited to subsequent generations was unknown at the time, until DNA was identified as the molecule for storing genetic information (Watson and Crick 1953).

Critical for the evolution of organisms is the erroneousness of DNA replication. Throughout time, cells divide, and with each cell division all DNA in the cell is replicated. DNA replication is not perfect, and this imperfection leads to random mutations in the DNA of a species over time (Kunkel 2004). These random mutations in the DNA can, through transcription into mRNA, be translated into changes in the amino acid sequence of a protein (Crick 1970), thereby affecting the organisms phenotype. In this way, random DNA mutations cause phenotypic variation within populations of species, on which natural selection can act. It is through this process of random DNA mutations and subsequent natural selection that species evolve over time.

1.2 The evolution of complex traits

The evolution of phenotypic traits through DNA mutations is generally quite intuitive for simple traits, for which one or a few mutations are sufficient to cause the observed phenotypic change. However, this process is somewhat less self-explanatory when imagining the evolution of complex traits, such as organs. Complex organs consist of multiple interlocking parts, each of them necessary for the correct functioning of the organ. For example, the human eye must include a cornea, lens, and retina to function properly. It is hard to imagine that each of these parts evolved separately, without any selective advantage before coming together in a fully functional organ. Therefore, it has been proposed that these organs were not complex at first, but gradually evolved from intermediate organ-like structures with related, but simpler functions (Dial 2003; Oakley and Speiser 2015).

The best studied example of complex trait evolution is that of the evolution of the eye. In the animal kingdom, eye complexity varies hugely between different clades, with functions ranging from simple light detection in ctenophores (comb jellies) to complex camera-like eyes capable of full image and depth vision in vertebrates, insects and some mollusks (Lamb et al. 2007; Nilsson 2013). These observations give an example for how these complex traits can evolve: simpler versions of the organs we observe today can already give a selective advantage (for instance simple light detection to allow for a circadian rhythm), allow for fixation of the mutations necessary for this trait, and provide a stable basis for the evolution of more complex varieties of the trait.

1.3 Comparative genomics as a means to study complex trait evolution

1.3.1 The principle of comparative genomics

Although phenotypic observations can give us insight into the stepwise manner in which complex traits evolve, this does not tell us anything about the genomic mutations causing the evolution of these traits. To find the genomic basis of complex trait evolution, the comparative genomics approach has been proposed to be effective (Hardison 2003; Miller et al. 2004). The principle of this approach is simple: if two animal species differ in their phenotype by a certain trait, this difference should be encoded in the genome. By sequencing and comparing the genomes of these two species, we should be able to obtain a list of genomic differences between the two species, which includes the differences that cause the difference in phenotype. The challenge here is to separate genomic differences that contribute to the phenotypic difference of interest from those that do not.

1.3.2 Methods in comparative genomics

The field of comparative genomics as we know it today is young compared to other fields in biology, as fully sequenced and assembled genomes were relatively rare until about five years ago. However, due to a steady decrease in sequencing costs, generating new genome assemblies has come within reach for individual research groups, sparking an interest for methods to compare these genomes with each other. Most well-known methods fall into one of three categories: comparing gene evolution, comparing evolutionary constraint, or finding structural variations.

Several methods exist to compare gene evolution in different genomes. One common method to infer patterns of selection impacting gene evolution is by estimating the ratio between the number of non-synonymous and synonymous nucleotide substitutions (d_N/d_S) for a coding sequence within the branches of a phylogeny. For a neutrally evolving sequence, there would be no effect of a non-synonymous mutation, and the d_N/d_S ratio would approach 1. Protein-coding genes however, are highly conserved, so purifying selection against non-synonymous mutations is expected ($d_N/d_S << 1$). Indeed, on average, protein-coding genes have a d_N/d_S ratio far below 1 (Studer et al. 2008). However, certain situations can favor synonymous changes in a protein sequence, for instance when a protein acquires a new function. This phenomenon is called positive selection and can lead to elevated d_N/d_S ratios at specific sites in the sequence, or branches in the phylogeny. Approaches to detect elevated d_N/d_S ratios have been applied in numerous studies, some examples include detecting positively selected pathways in primate lineages (Daub et al. 2017), scanning for different evolutionary adaptations between two algae species (Teng et al. 2017), or detecting positive selection in plastid genes after a change in a photosynthetic mechanism (Piot et al. 2018).

Another method for finding deviations from the expected genomic evolution of certain genes is the so-called evolutionary rate analysis. For this method, the mutation rate of certain genes is compared between two groups: the species that have the phenotypic trait of interest, and species that do not. Consistent deviations in evolutionary rate between the two groups may indicate an association of this gene with the observed phenotypic evolution. This approach has previously been successfully applied for the evolution of aquatic adaptations in marine mammals (Chikina et al. 2016), as well as for adaptations in subterranean mammals (Partha et al. 2017), but requires multiple independent evolutions of the same or a similar trait, a phenomenon that is relatively rare for complex traits.

To compare the genomes of species outside genic regions, an evolutionary constraint analysis is often performed to find genomic regions that are functional, but do not code for any proteins. The principle behind this analysis is that if a genomic region has a crucial function for the organisms involved, the mutation rate in this region will be lower than that of neutrally evolving regions because of the possible detrimental effect of these mutations. Therefore, genomic regions can be classified into functional and non-functional regions by comparing mutations rates between species across the genome. Examples of tools applying this principle include PhastCons, GERP++ and SiPhy (Zhang et al. 2008; Garber et al. 2009; Davydov et al. 2010). The genomic distribution of these conserved regions can be compared between species that differ in the phenotypic trait of interest, which can give insights in the regulatory changes that occur when the trait evolves.

The aforementioned methods have in common that they are designed to function on single-copy regions of the genome. However, structural variants such as gene duplications are often associated with adaptive evolution as well, and detecting these variants is essential to get a complete picture of genome evolution (Prud'homme et al. 2007; Wagner 2008). Finding duplicated regions when considering only two genomes of related species is usually quite straightforward using a pairwise genome aligner such as MUMmer (Marçais et al. 2018). Genomic regions were two regions of one genome align to one region of the other genome can be extracted as candidate regions and investigated further for gene duplications. When considering larger groups of genomes, usually species pairs with an outgroup are constructed to reduce the complexity of the analysis somewhat, as pairwise comparisons against an outgroup are easier to interpret than all-versus-all comparisons. An example of an application of this method can be found in comparing freshwater with saltwater populations of the three-spined stickleback (Hirase et al. 2014). Here, a consistent association was found between gene duplications and adaptation to a freshwater environment. Another example can be found in Antarctic notothenioid fish, where gene copy numbers of genes related to cold adaptation had drastically increased compared to their temperate counterparts (Chen et al. 2008).

1.3.3 Factors to consider when applying a comparative genomics method

Regardless of the method used, several factors are important in maximizing the statistical power to find genomic changes associated with phenotypic change. First, multiple independent evolutionary origins of the trait of interest are essential to gain statistical power: assuming that the genomic basis for the different evolutionary origins is similar, the set of candidate genes can be narrowed down to those that show a deviation from the null hypothesis in all of those evolutionary origins.

Second, comparing closely related species is generally more effective than comparing more distant relatives. Assuming that the evolution of a complex trait is accompanied by a roughly proportionally complex genomic change, the magnitude of this change will stand out more against the background of neutral genomic mutations when this complex trait evolves in a relatively short amount of time. Therefore, comparing closely related species that still differ in the trait of interest will increase the signal-to-noise ratio of any analysis that aims to find genomic mutations associated with phenotypic change.

Third, high quality genome assemblies are essential. Comparing two genome assemblies of the same species, one of which has been made ten years ago and another one that has been made this year will yield a large list of, mainly structural, differences. These differences rarely represent biological differences, but rather the technical limitations of the sequencing techniques involved. It is therefore important to be aware of the effect of sequencing technique on the resulting genome assembly. This is especially true for analyses that involve structural variant detection, as these types of variants are most affected by genome assembly quality (Mahmoud et al. 2019).

Fourth, although not technically important for maximizing statistical power, including species showing intermediate stages of complexity of the trait of interest can yield additional insight, as these species are expected to display an “intermediate” genotype as well, with some but not all causal mutations being present in their genomes. This can also reveal information about the directionality by which these complex traits evolve on a genomic level: if certain key mutations are necessary for the evolution of a complex trait, these should also be found in species showing intermediate stages of complexity.

In short, to optimally study the genomics of complex trait evolution we would need a model in which the complex trait has evolved several times in closely related species, with intermediate stages of complexity still being present in present-day species. Additionally, practical availability of biological material is essential for the generation of high-quality genome assemblies. Such a model can be found in the livebearing fish family Poeciliidae.

1.4 The poeciliid placenta: an excellent model to study complex trait evolution

The Poeciliidae (*sensu* (Parenti 1981)) are a family consisting of around 275 species of small livebearing fish, living in both South and Central America (Reis et al. 2003; Van Der Laan et al. 2014). All species in this family except for one (the egg-laying *Tomeurus gracilis* (Parenti et al. 2010)) bear live young, but the mode of nutrient provisioning to the offspring differs per species. The ancestral state of nutrient provisioning to the offspring in the Poeciliidae is thought to be lecithotrophy

(Furness et al. 2019), which is the feeding of the offspring via egg yolk proteins. In this mode, nutrients are supplied to the egg cell before fertilization. However, in several species a deviation from the ancestral state has been observed, with the majority of nutrients being supplied after fertilization through a placenta (Turner 1940; Pollux et al. 2009). This mode of provisioning is called matrotrophy.

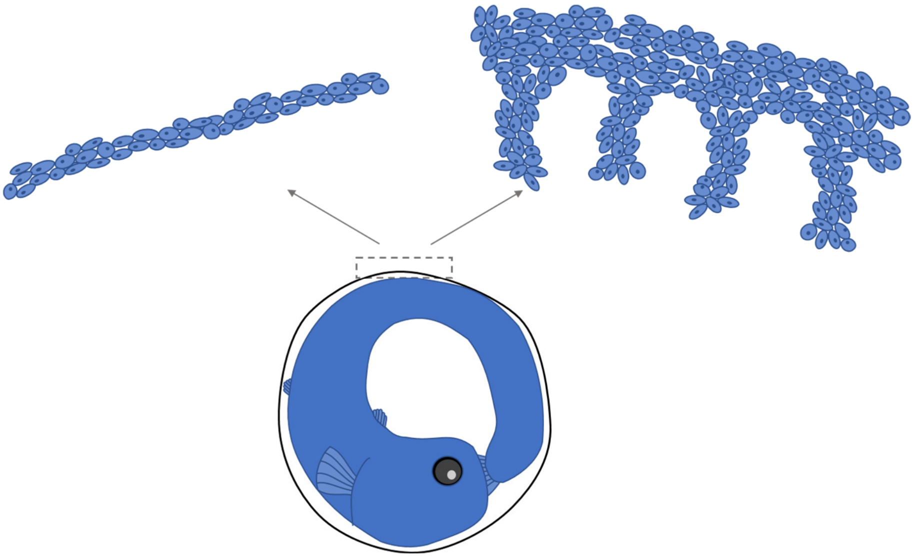

Figure 1.1 Schematic overview of the poeciliid follicular placenta. Individual embryos of poeciliid species grow within a follicle (bottom). In lecithotrophic species, the follicular wall consists of a thin layer of cells (top left). In some extensive matrotrophic species, the follicular wall is thick and extensively folded (top right).

The complexity of the placenta of poeciliid species is often approximated using the Matrotrophy Index (MI), which is defined as the weight of the offspring at birth divided by the weight of the egg at fertilization (Wourms et al. 1988). Lecithotrophic species provide all nutrients before fertilization, which means the offspring at birth will have lost weight relative to the egg at fertilization due to metabolic costs, leading to a MI lesser than 1. In matrotrophic species the embryo weight will increase throughout pregnancy due to nutrients being supplied via the placenta, leading to a MI greater than 1. MI values in the Poeciliidae range from 0.6 for fully lecithotrophic species to 117 for the highly matrotrophic *Poeciliopsis retropinna* (Pollux et al. 2014). Morphological studies show that these MI values correlate well with complexity of the placental tissue of the species: both the thickness and the degree of folding of the follicular tissue surrounding the embryo correlates well with MI (Wourms et al.

1988; Grove and Wourms 1991; Grove and Wourms 1994; Kwan et al. 2015). Therefore, in this thesis, MI will be used as a proxy for placental complexity in poeciliid fish.

As a model for complex trait evolution, the Poeciliidae meet all aforementioned requirements. First, phylogenetic analysis has shown that the species that supply nutrients through a placenta are not monophyletic within this family, instead it has been estimated that the poeciliid placenta has evolved nine times independently (Furness et al. 2019). This allows for comparison of different instances of placenta evolution and, assuming genomic convergence, for a high statistical power in finding the genes involved in placenta evolution. Second, the most recent common ancestor of the Poeciliidae has been estimated to have lived around 54 million years ago (Reznick et al. 2017), which is relatively recent for the evolution of a new organ. As a comparison, the mammalian placenta is estimated to be 200 to 250 million years old (Tarver et al. 2016). Additionally, most of the poeciliid placentas have evolved much more recently than 54 million years ago. For instance, the three instances of placenta evolution in the genus *Poeciliopsis* have been estimated to be between 2 and 10 million years old, and have evolved within between 0.75 and 2.36 million years (Reznick et al. 2002; Reznick et al. 2017). This more recent evolution means that causal changes in the genome will be easier to find, as the genomes have changed to a lesser extent since a placenta has evolved than would be the case for, for instance, placental mammals. Third, intermediate stages of placental complexity can be found within this family as well, as some species provide a part of the nutrients before fertilization, and another part after fertilization. Examples of this can be found in the genus *Gambusia*, as well as the subgenus *Mollinesia* (Trexler 1985; Marsh-Matthews et al. 2001; DeMarais and Oldis 2005). This suggests that placenta evolution is an ongoing process in the Poeciliidae, and that genes involved in placenta evolution may be under the influence of natural selection, allowing us to detect them with established evolutionary analyses.

1.5 Hypotheses on poeciliid placenta evolution

As the poeciliid placenta has evolved nine times independently, it seems self-evident that having a placenta provides a certain selective advantage in poeciliid fish, at least under some circumstances. Several hypotheses have been put forward that try to explain the repeated emergence of this organ in poeciliid fish. Three of these hypotheses are the most well-known: the locomotor cost hypothesis, the resource-availability hypothesis, and the viviparity-driven conflict hypothesis.

The locomotor cost hypothesis centers around the idea that pregnancy reduces locomotory performance in female fish, because pregnant females are heavier and less streamlined compared to their non-pregnant counterparts. This makes them more vulnerable to predators (Ghalambor et al. 2004). For non-placental species, this burden is on average heavier than for placental species, because they have to produce much bigger eggs as they have to contain all necessary resources for the fetus at once. Placental species can start with smaller eggs, which do not have such a big effect on locomotory performance, and feed their embryos throughout pregnancy. This implies that placental species have a selective advantage compared to their non-placental counterparts in habitats with a high predation pressure or water flow velocity. Indeed, evidence was found supporting this theory in populations of *Poeciliopsis turrubarensis*, as well as *Poeciliopsis turneri* (Jaime Zúñiga-Vega et al. 2007; Fleuren et al. 2018; Hagmayer et al. 2020).

The resource-availability hypothesis was originally proposed by Trexler, suggesting that matrotrophic species can have a higher reproduction rate compared to lecithotrophic species in situations of high food abundance, because they can spread out their resource allocation over the whole course of their pregnancy (Trexler 1997). This would mean that in habitats with high and constant food availability, matrotrophic species would have a selective advantage over lecithotrophic species. Indeed, there seems to be some evidence that higher food abundance favors matrotrophy over lecithotrophy as a reproduction strategy (Pollux and Reznick 2011; Riesch et al. 2012), although some conflicting evidence also exists (Marsh-Matthews and Deaton 2006; Banet and Reznick 2008).

The viviparity-driven conflict hypothesis proposes that there is a conflict of interest between the mother and her offspring over the amount of resources allocated by the mother to her offspring. For the mother, it is optimal to provide sufficient resources to as many offspring as possible, as every fetus carries the genetic material of the mother in an equal degree. However, for the offspring it is more beneficial to demand as much resources as possible, thereby increasing its own chance of survival (Trivers 1974; Zeh and Zeh 2000). This conflict would then cause an evolutionary "arms race" between mother and offspring at the site of nutrient transfer: the placenta. Consequently, this would lead to accelerated evolution of the placenta, leading to a diverse variety in degree of placentation as observed in *Poeciliids*. Direct evidence for accelerated evolution of placenta associated genes in livebearing fish is scarce, although one study provides evidence for accelerated evolution of the *igf2* gene in poeciliid fish (O'Neill et al. 2007). Indirect evidence of parent-offspring conflict can be found in so-called genomic imprinting, which is the mono-allelic expression of genes based on parent-specific methylation patterns (Li et al. 1993a; Reik and Walter 2001). Genes associated with neonatal growth are often found to be imprinted in mammals: growth promoting genes like Igf2, Peg1, Peg3 are often expressed only from the paternal allele, while growth repressors are maternally expressed (Barlow and Bartolomei 2014).This suggests that genomic imprinting is used as a “weapon” in the parent-offspring conflict, at least in mammals. However, in livebearing fish evidence for this has not been found. In fact, the *igf2* gene was found to be expressed bi-allelically in the embryos of *Poeciliopsis prolifica* and *Heterandria formosa* (Lawton et al. 2005).

To summarize, several theories exist on the evolution of the placenta in poeciliid fish, none of which are mutually exclusive. Although it is unclear which of these predicted processes influence the occurrence of placenta evolution the most, it is clear that the evolution of the placenta is an adaptation to certain environments. This adaptive evolution should leave its marks on the genomic level as well.

1.6 Thesis outline

In this thesis, I aim to find the genomic basis of the repeated evolution of the placenta in the livebearing fish family Poeciliidae. This process consists of two steps: first, I will sequence and assemble the genome of multiple poeciliid species. Although several genome assemblies of poeciliid species were already available before the start of this thesis project, all of these species are non-placental (Schartl et al. 2013; Künstner et al. 2016; Shen et al. 2016b). Therefore, my sequencing efforts will focus on placental species, together with their closest non-placental relatives. Second, I will compare the genomes of placental species to the genomes of non-placental species in an effort to find genomic differences that occur consistently in association with the evolution of the poeciliid placenta.

The first two research chapters focus on the sequencing and assembly of new genomes. In **chapter 2**, I re-assemble the genome of the placental *Heterandria formosa* using existing sequencing data, and look for genes that exhibit signs of positive selection, while these signs are absent in the non-placental poeciliids *Poecilia reticulata*, *Poecilia formosa* and *Xiphophorus maculatus*. In **chapter 3**, I sequence and assemble the genomes of the placental *Poeciliopsis retropinna* and its non-placental relative *Poeciliopsis turrubarensis* using third-generation sequencing techniques. With the resulting assemblies, I assess gene duplications and deletions that occur in the placental *P. retropinna*, but not in the non-placental *P. turrubarensis*.

The next two research chapters focus on finding genomic changes associated with placenta evolution. In **chapter 4**, I compare the genomes of 26 poeciliid species, 15 of which were assembled for this study. I focus on the rate of evolution of orthologous genes throughout this family, attempting to associate this rate with placenta evolution. Additionally, I look for consistent differences in the presence or absence of conserved non-coding elements in the genomes of placental species *versus* the genomes of non-placental species. In **chapter 5** I use 12 high-quality poeciliid genomes to test the hypothesis that the poeciliid placenta evolution is consistently associated with certain structural variants, such as gene duplications or deletions.

The first five chapters are about linking genomic variation to phenotypic observations. However, phenotypic variation can also be caused through epigenetic means, such as methylation. In **chapter 6** I develop and apply a new method to detect allele-specific DNA methylation in *Poeciliopsis gracilis*, and investigate how these instances of allele-specific methylation are inherited from parent to offspring. Finally, in **chapter 7** I discuss my findings and provide recommendations for future research.

2

The genome of the livebearing fish *Heterandria formosa* implicates a role of conserved vertebrate genes in the evolution of placental fish

Henri van Kruistum[1,2], Joost van den Heuvel[3], Joseph Travis[4], Ken Kraaijeveld[5,6] Bas J. Zwaan[3], Martien A.M. Groenen[1], Hendrik-Jan Megens[1,7] and Bart J.A. Pollux[2]

[1] Animal Breeding and Genomics, Wageningen University, The Netherlands. [2]Experimental Zoology, Wageningen University, The Netherlands. [3]Laboratory of Genetics, Wageningen University, The Netherlands. [4]Department of Biological Science, Florida State University, USA. [5]Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, The Netherlands. [6]Leiden Genome Technology Center Department of Human Genetics, Leiden University Medical Center, The Netherlands. [7]Aquaculture and Fisheries Group, Wageningen University, The Netherlands.

BMC Evolutionary biology (2019), 19(1), 1-11

Abstract

The evolution of complex organs is thought to occur via a stepwise process, each subsequent step increasing the organ's complexity by a tiny amount. This evolutionary process can be studied by comparing closely related species that vary in the presence or absence of their organs. This is the case for the placenta in the live-bearing fish family Poeciliidae, as members of this family vary markedly in their ability to supply nutrients to their offspring via a placenta. Here, we investigate the genomic basis underlying this phenotypic variation in *Heterandria formosa*, a poeciliid fish with a highly complex placenta. We compare this genome to three published reference genomes of non-placental poeciliid fish to gain insight in which genes may have played a role in the evolution of the placenta in the Poeciliidae.

We sequenced the genome of *H. formosa*, providing the first whole genome sequence for a placental poeciliid. We looked for signatures of adaptive evolution by comparing its gene sequences to those of three non-placental live-bearing relatives. Using comparative evolutionary analyses, we found 17 genes that were positively selected exclusively in *H. formosa*, as well as five gene duplications exclusive to *H. formosa*. Eight of the genes evolving under positive selection in *H. formosa* have a placental function in mammals, most notably endometrial tissue remodeling or endometrial cell proliferation.

Our results show that a substantial portion of positively selected genes have a function that correlates well with the morphological changes that form the placenta of *H. formosa*, compared to the corresponding tissue in non-placental poeciliids. These functions are mainly endometrial tissue remodeling and endometrial cell proliferation. Therefore, we hypothesize that natural selection acting on genes involved in these functions plays a key role in the evolution of the placenta in *H. formosa*.

Key words: *Heterandria formosa*, Poeciliidae, placenta, matrotrophy, positive selection, gene duplication, molecular evolution, whole genome sequencing

2.1 Background

Explaining the evolution of complex organs, consisting of multiple interacting parts, is one of the greatest challenges in evolution. Charles Darwin was the first to propose an explanation for this phenomenon; in his seminal work on natural selection, he hypothesized that complex organs were not complex at first, but gradually evolved into what we observe today (Darwin 1859). However, finding examples of this stepwise process poses a challenge, mainly because of two reasons. First, species possessing an organ of intermediate complexity have often gone extinct, leaving the present-day observer with only the end-result of a long series of potentially minute evolutionary steps. Second, when differences in organ complexity between species exist, these species are often separated by a large phylogenetic distance, sharing only a very remote common ancestor. For instance, intermediate stages of complexity can be found in the mollusk eye (Ekström and Meissl 2003; Fernald 2006). However, the different types of mollusk eyes are found in distantly related taxa, which diverged about half a billion years ago. This makes a comparative analysis on a genomic level not straightforward. To truly understand how molecular pathways are altered during evolution to give rise to complex organs, a model system is required that has recently evolved a complex organ with the ancestral and intermediate states still extant in closely related species. Ideally, such a complex organ should have originated multiple times, e.g. due to convergent evolution resulting from similar evolutionary pressure. Such a model system can be found in the development of the placenta in the livebearing fish family Poeciliidae (Reznick et al. 2002).

The placenta is an organ that facilitates nutrient exchange between mother and offspring. It is present in all major vertebrate lineages, although its anatomical details differ between taxa (Blackburn 2015; Griffith and Wagner 2017). Numerous genes involved in placental development have been identified, making the placenta a prime example of complexity (Rossant and Cross 2001; Cross et al. 2003; Hou et al. 2009). Most research on the placenta has been performed in eutherian mammals. Eutherian mammals, however, are limited in their suitability to study the evolution of the placenta, because all contemporary placental mammals (i) inherited their placenta from a single common ancestor that lived >160 million years ago, and (ii) all have complex placentas and have no close living relatives that lack placentas. By contrast, the placenta has been estimated to evolve independently nine times in amphibians, and 12 times in ray-finned fish (Blackburn 2015).

There are three reasons to focus on placental evolution of the live-bearing fish family Poeciliidae. First, the placenta has evolved independently at least eight times in the Poeciliidae (Pollux et al. 2009). This makes it possible to compare different instances of placental evolution within closely related species. Second, intermediate stages of placental complexity exist within this family. In fact, placental complexity in the Poeciliidae seems to vary continuously amongst species, rather than species either having a placenta or not (Reznick et al. 2002). Third, all of this variation is present among relatively closely related species. This allows us to more easily compare the genomes of these species. A genomic comparison between species varying in placental complexity may unveil the genomic basis underlying this difference in complexity.

The degree of maternal provisioning in the family Poeciliidae has been quantified in the Matrotrophy Index (MI), which is the estimated dry mass of offspring at birth divided by the dry mass of the egg at fertilization (Wourms et al. 1988). Poeciliid fish have a MI ranging from 0.6 for non-placental (lecithotrophic) species to more than 100 for species with a highly complex placenta (matrotrophic), with species exhibiting intermediate values also being present (Reznick et al. 2002). The MI can act as a proxy for placental complexity, because species with a high MI have a more complex placenta compared to species with a low or intermediate MI (Turner 1940; Grove and Wourms 1994; Kwan et al. 2015; Olivera-Tlahuel et al. 2018). The main differences lie in the structure of the maternal follicular epithelium. The unspecialized follicular wall of lecithotrophic (non-placental, MI < 1) species is very thin and plays no role in maternal provisioning (Turner 1940; Jollie and Jollie 1964). In matrotrophic (placental) species the follicular epithelium is much thicker, more extensively folded and features specialized adaptations that facilitate maternal-to-embryo nutrient transfer, such as a high vascularization, a high density of microvilli, and the presence of specialized cytoplasmic organelles (Grove and Wourms 1994; Kwan et al. 2015). Given the co-occurrence of these structural tissue features with a high MI, it is likely that these adaptations facilitate extensive matrotrophy.

Early studies on natural selection at the molecular level in the family Poeciliidae have compared genes of one or more poeciliid species to genes of other more distantly-related teleosts (Schartl et al. 2013; Jue et al. 2018; Warren et al. 2018), or the analysis was limited to one or only a few genes known to be involved in placenta development in mammals (O'Neill et al. 2007; Schartl et al. 2013). Exhaustively identifying genes responsible for placentation is impossible in such approaches, because large differences in placental complexity exist *within* the family Poeciliidae. In the present study, therefore, natural selection is investigated between more closely related species, focusing on the genomic differences between lecithotrophic and matrotrophic species within the family Poeciliidae.

Here, we investigate the genomic basis of placental complexity by exploring the genome of a highly matrotrophic poeciliid: the least killifish, *Heterandria formosa*. This species has a MI of around 35, and morphological analysis has shown that it has a highly complex placenta (Grove and Wourms 1994). Specifically, we aim to, (1) sequence the genome of *H. formosa*, providing the first whole genome sequence of a matrotrophic poeciliid, and (2) compare this genome to published reference genomes of three related lecithotrophic species: the Trinidadian guppy (*Poecilia reticulata*) (Künstner et al. 2016), the Amazon molly (*Poecilia formosa*) (Warren et al. 2018), and the Platyfish (*Xiphophorus maculatus*) (Schartl et al. 2013). These latter three species are lecithotrophic (MI < 1), and lack a placenta. Such large difference in placentation in closely related species may suggest the involvement of natural selection, which should be visible in associated signatures of selection in the genome. Comparing genes evolving under positive selection to their orthologs in three non-placental species allows prioritization of genes related to placentation; genes showing evidence of positive selection in *H. formosa*, but not in any of its lecithotrophic relatives are likely enriched for involvement in placentation. Additionally, we identified genes that have likely been duplicated in the genome of *H. formosa*, using a combination of breakpoint and read-depth based methods. Gene duplications are known to be an important driving force of adaptive evolution, so it is plausible that an increased placental complexity is associated with distinct gene duplications (Lynch 2002). Through these methods we identify a number of genes that have likely contributed to phenotypic variation in, and evolution of, placentation in the family Poeciliidae.

2.2 Results

2.2.1 Whole genome sequencing of *Heterandria formosa*

We sequenced the genome of *H. formosa* to an average coverage of 40X, yielding 90 Gb data containing 182 million 150 bp paired-end reads. The genome was assembled using SPAdes assembler (Bankevich et al. 2012), resulting in a draft assembly with a size of 722 Mb. *H. formosa* genome size estimation based on k-mer analysis showed an estimated genome size of 670 Mb, which is slightly lower than the assembly size. This is possibly a result of the relatively high heterozygosity of the sample leading to redundant contigs, as the sequenced individual was not from an inbred population. To reduce this redundancy, redundans (Pryszcz and Gabaldón 2016) was run on the assembly to remove heterozygous contigs, and rescaffold the assembly based on paired-read information. This reduced the assembly size to 608 Mb, which is slightly lower than the estimated genome size, and also lower than other poeciliid genome assemblies (Schartl et al. 2013; Künstner et al. 2016; Warren et al. 2018). Additionally, scaffold N50 increased from 11 Kb to 26.5 Kb by the rescaffolding procedure. The lower assembly size compared to the estimated genome size can be explained by the fact that this assembly was based on short reads, and some repetitive sequences will likely be collapsed in the assembly, leading to a somewhat smaller assembly size. Summary statistics of this genome assembly are listed in table 2.1.

Table 2.1 Summary statistics for the *H. formosa* genome assembly

Assembly size	608 Mb
Contig N50	6108 bp
Largest contig	77373 bp
Scaffold N50	26563 bp
Largest scaffold	226934 bp
GC content	38.59%
Heterozygosity	1 in 203 sites

The genome of *H. formosa* was aligned to the reference genome of *P. reticulata* using LAST (Kiełbasa et al. 2011) . The majority of the scaffolds of the *H. formosa* assembly aligned to one linkage group in *P. reticulata* (figure 2.1B), suggesting extensive synteny between the two species. For some smaller contigs, no match to *P. reticulata* linkage groups was found (Figure 2.1A). All *P. reticulata* linkage groups were covered roughly equally by the *H. formosa* contigs, covering around 80% of the bases in *P. reticulata* (figure 2.1C). This means that around 20% of the *P. reticulata* bases were not covered by any *H. formosa* sequence, which may be because the *H. formosa* assembly is smaller than the *P. reticulata* assembly, or because there is high sequence divergence in these regions.

The coverage drops at the edges of the linkage groups, likely reflecting the underrepresentation of repetitive sequences in the *H. formosa* assembly due to it being assembled from paired-end reads only. A portion of the *H. formosa* contigs (23% base fraction) was split in the alignment to two *P. reticulata* linkage groups (figure 2.1B). For a minority (10%) of these contigs, alignment length was longer than 1000 bp for both linkage groups to which the contig aligned. This observation suggests that some genomic rearrangements may have occurred. For contigs aligning to three or more linkage groups, alignments were generally very short for all but one linkage group, indicating that this is most likely a result either of contigs aligning to ambiguous regions in the genome, or assembly errors.

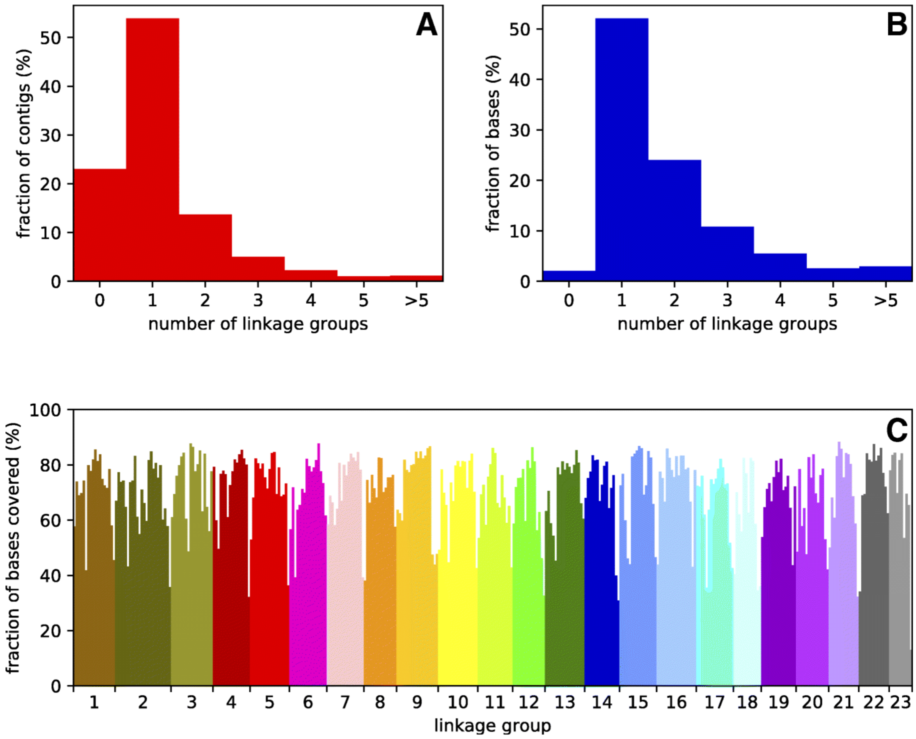

Figure 2.1 A: Fraction of H. formosa assembly contigs aligning to a certain number of linkage groups of the P. reticulata genome assembly. **B**: Fraction of bases in H. formosa contigs that align to a certain number of linkage groups of the P. reticulata genome assembly. **C**: Percentage of P. reticulata bases covered by the 1:1 H. formosa:P.reticulata alignment, 2Mb bins.

2.2.2 Positive selection

We identified 8,056 1:1:1 gene orthologs between *P. reticulata*, *P. formosa* and *X. maculatus* using ProteinOrtho (Lechner et al. 2011). From these genes, we retrieved the complete coding sequences of 6,774 genes in *H. formosa* through the whole genome alignment with *P. reticulata*. Using the codeml program of the PAML package (Yang 2007), we tested these genes for both positive selection across all investigated poeciliid species, and positive selection in *H. formosa*. At 10% FDR, we found 104 genes to be positively selected across the whole phylogeny and 29 genes to be positively selected in *H. formosa*. Eleven genes were significant for both tests, leaving 18 genes exclusively positively selected in the *H. formosa* lineage (table 2.2). In one case, a stop-gained mutation was observed inside the first exon, so this protein was left out of the final results.

Table 2.2 positively selected genes in *H. formosa* (10% FDR).

Gene symbol	gene name	p-value
pla2g2a	Phospholipase A2 Group IIA	3.32E-07
timp4	Tissue Inhibitor Of Metalloproteinases 4	2.92E-06
rbl1	Retinoblastoma-Like 1	1.63E-05
cldnd	Claudin d	2.28E-05
tmem230	Transmembrane Protein 230	3.34E-05
kiaa1324/eig121	Estrogen Induced Gene 121	3.43E-05
pnkd	Paroxysmal Nonkinesigenic Dyskinesia	6.69E-05
mmp15	Matrix metalloproteinase 15	2.50E-04
gpr34	G Protein-Coupled Receptor 34	2.55E-04
btbd7	BTB Domain Containing 7	2.68E-04
glp1	Glucagon-like peptide 1	2.84E-04
cldn4	Claudin 4	3.04E-04
slc35d3	Solute Carrier 35 Member d3	3.27E-04
pcdh10	Protocadherin-10	4.05E-04
loc103465290	Uncharacterized protein	4.51E-04
allc	Allantoicase	4.58E-04
slc20a1	Solute Carrier Family 20 Member a1	5.60E-04

A substantial number of these genes have placental functions in mammals. First, *pla2g2a* was isolated from human placenta (Buhl et al. 1995), and evidence found in horse points to a function in placental steroid metabolism (Ababneh and Troedsson 2013). However, activity of this protein is not limited to placenta, and has been linked to the immune system as well (Saegusa et al. 2008). Second, a matrix metalloproteinase and a matrix metalloproteinase inhibitor (*mmp15* and *timp4*) were both positively selected in *H. formosa*. Both proteins are involved in endometrial tissue remodeling and placental labyrinth formation (Yang et al. 2006; Szabova et al. 2010). Third, *rbl1* and *kiaa1324* gene expression has been linked to endometrial cell proliferation (Cavallotti et al. 2001; Deng et al. 2010). Fourth, genes coding for two claudin proteins (*cldnd* and *cldn4*) were found to be positively selected in this analysis. Claudins are cell-cell adhesion proteins known to be essential in placental tight junctions, regulating ion transport (Aplin et al. 2009; Ahn et al. 2015). Interestingly, claudins are also involved in tissue remodeling by interacting with matrix metalloproteinases (Miyamori et al. 2001; Gaetje et al. 2008). Finally, *btbd7* is involved in tissue remodeling of embryonic epithelial cells by interacting with cell-cell adhesion proteins (Onodera et al. 2010), and is associated with preeclampsia in humans (Jia et al. 2012). We searched for expression of these genes in the human protein atlas (Uhlén et al. 2015) and the tissue-specific transcriptome of the closely related *Poeciliopsis prolifica* (Jue et al. 2018). All of these proteins are expressed in the human placenta, except for k*iaa1324,* which is more active in the endometrium (supplementary table 2.1). In *P. prolifica*, we found expression of all of these genes in either placental or ovarian tissue, except for *pla2g2a* (supplementary table 2.1).

As for the remaining nine positively selected genes in *H. formosa*, most are neuron associated (*pnkd, tmem230, pcdh10, gpr34, slc35d3*) (Wolverton and Lalande 2001; Shen et al. 2015; Deng et al. 2016; Wei et al. 2016; Zheng et al. 2017), which suggests ongoing selection on behavioral traits as observed earlier in poeciliids and teleost fish in general (Bisazza 1993; Bshary et al. 2002). The four remaining genes evolving under positive selection in *H. formosa* have varying or unknown functions. For a further elaboration on all genes found to be evolving under positive selection in *H. formosa*, see supplementary table 2.1.

To assess the function of positively selected genes in a quantitative manner, GO term enrichment analysis was performed using GOrilla (Eden et al. 2009). The enriched GO terms with the lowest p-value were associated with cell-cell adhesion. Other enriched GO terms of interest were negative regulation of endopeptidase activity, dopamine and catecholamine metabolism, positive regulation of cytosolic calcium ion concentration, and cell migration. For all results of the GO enrichment analysis, see supplementary table 2.2.

The evolution of complex structures may also involve changes in gene function and to investigate this possibility in *H. formosa*, we employed Bayes Empirical Bayes (BEB) analysis with PAML to infer which codons in the coding sequence are most likely subject to positive selection and thereby obtain information about a possible change of function. Two examples of this inference are shown for the *Timp4* and *mmp15* genes (figure 2.2 and 2.3). As shown in the figure, positive selection in *H. formosa* Timp4 is widespread throughout the protein, as 20 out of 224 codons are predicted to be under positive selection (p > 80%). Positively selected sites interfere with residues of both the metzincin- as well as the hemopexin-binding domain, although most residues of these domains remain conserved. This may indicate a change in function, for instance in the type of metalloproteinases the protein binds to. Positively selected sites in Mmp15 are located next to and in between the catalytic and hemopexin (metal binding) domains, but do not overlap with the active residues. Little is known about these regions of the protein, but its catalytic function is not likely affected.

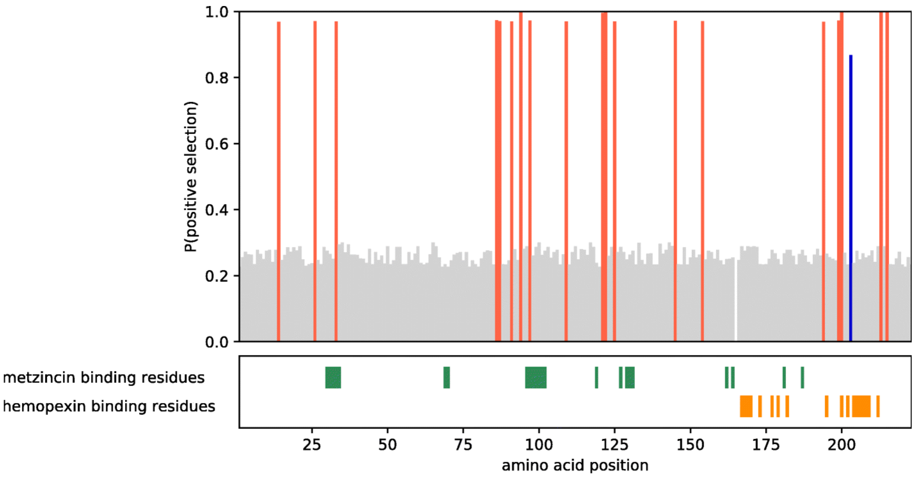

Figure 2.2 Likelihood of positive selection for each codon in H. formosa Timp4. Active residues are plotted on the bottom panel. Color codes for probability of positive selection: Red > 95% > blue > 80% > grey.

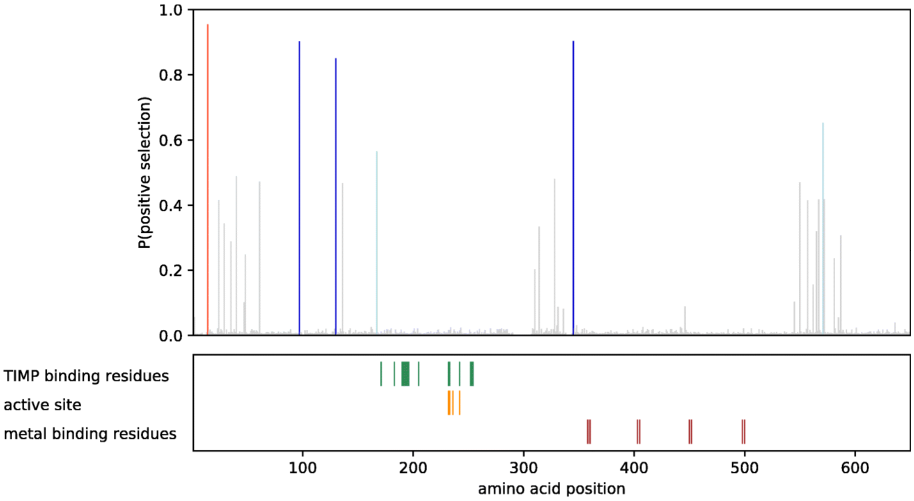

Figure 2.3 Likelihood of positive selection for each codon in H. formosa Mmp15. Active residues are plotted on the bottom panel. Color codes for probability of positive selection: Red > 95% > blue > 80% > light blue > 50% > grey.

2.2.3 Gene duplications

Potential gene duplications were identified by mapping reads from *H. formosa* to the P. reticulata genome, and identifying potential breakpoints by running Lumpy (Layer et al. 2014) on the alignment. Combining the breakpoints with a read depth signal allowed for identifying potential duplications. Using this method, we identified 46 potentially duplicated segments. However, after manual evaluation (see methods) only six of these segments were retained as likely true duplications, reflecting the difficulty to identify true duplications using short reads. These segments are given in table 3.

Table 3. Duplicated regions in *H. formosa*

Duplicated area (position on *P. reticulata* genome)	Length (bp)	Genes
NC_024349.1:9797934-9803865	5931	Overlaps with *Cdh1*
NC_024331.1:4200605-4202283	1678	None
NC_024335.1:30069829-30071128	1299	Overlaps uncharacterized protein
NC_024338.1:15595450-15598894	3444	Contains *Pla2g2a*
NC_024345.1:3814448-3869038	54590	Overlaps with *Camk2g*, *Ccdc88a*
NC_024333.1:20591856-20595250	3394	Contains *Urah*

Although not many genes were found to be duplicated in *H. formosa*, the results are concordant with the results from the positive selection analysis. Firstly, the gene coding for 5-hydroxyisourate hydrolase (*Urah*) was duplicated. This gene belongs to the same uric acid degradation pathway as the positively selected gene allantoicase (*Allc*). Secondly, *Pla2g2a*, which we showed above to be evolving under positive selection, is duplicated completely. Thirdly, a cadherin gene (*Cdh1*) appeared partially duplicated, in addition to *Pcdh10* evolving under positive selection. *Cdh1* expression is also known to be regulated by *Btbd7* (Daley et al. 2017), a gene found to be positively selected in *H. formosa*. Finally, a relatively large duplication containing the majority of the *Camk2g* gene and a small part of the *Ccdc88a* gene was observed. Both of these genes are involved in neural development (Rose et al. 2006).

2.3 Discussion

In this study, we sequenced and assembled the genome of *H. formosa*, a matrotrophic poeciliid. We aimed to use this information to gain insight in the evolution of the placenta in *H. formosa*, by looking for signatures of natural selection in its genome. One difficulty in identifying genes responsible for placentation is that natural selection in poeciliids is not limited to matrotrophy associated genes. For instance, immunity-related genes are consistently fast evolving in most vertebrate species, as a consequence of an evolutionary “arms race” between host immunity and pathogens (for instance (Boller and He 2009; Anderson et al. 2010)). Furthermore, it is known that courtship behavior is selected for in the family Poeciliidae (Bisazza 1993; Pollux et al. 2014), which implies that many genes associated with behavior are likely under the influence of sexual selection. These and other ongoing processes will cause coinciding genomic signatures of selection when considering selection acting on matrotrophy associated genes. We selected against these coinciding signatures of selection by distinguishing between positive selection across all investigated poeciliids and positive selection only observed in *H. formosa*, assuming that genes which are positively selected in both matrotrophic and lecithotrophic poeciliids are not likely responsible for the differences in placentation between the two groups.

Using this strategy, we identified 18 genes evolving under positive selection exclusively in *H. formosa*. Additionally, we identified six duplicated segments affecting a small number of genes. Significantly, mammalian orthologs of a substantial number of these genes are known to be involved in placental function and development, although most of these genes have different functions as well. For instance: protocadherin-10 (*Pcdh10*) is positively selected in *H. formosa*, and expressed in the human placenta (Wolverton and Lalande 2001). Cadherins are known to be important for placental cell-cell adhesion (Aplin et al. 2009). However, *Pcdh10* is also involved in certain parts of the brain associated with visual and olfactory function (Hirano et al. 1999), thus selective pressure on this gene could also occur because of selection on behavioral traits. Distinguishing between significance in placenta functioning or other functions was further evaluated by comparing gene function to the morphological differences between the placenta of *H. formosa* and that of its lecithotrophic relatives.

The main morphological differences in the placenta between matrotrophic and lecithotrophic poeciliids are found in the follicular epithelium, which is thicker and more extensively folded in matrotrophic species (Kwan et al. 2015). For a number of genes found to be positively selected in *H. formosa* it is possible they play a role in this change in tissue structure, most notably *mmp15* and *timp4*. Matrix metalloproteinases and their inhibitors are responsible for tissue remodeling (Lu et al. 2011), and both *mmp15* and *timp4* are active in the mammalian placenta (Yang et al. 2006; Szabova et al. 2010). Therefore, it is plausible that positive selection acting on these genes could result in a difference in placental morphology. Similarly, claudins are also involved in endometrial tissue remodeling, by activating matrix metalloproteinases (Miyamori et al. 2001; Gaetje et al. 2008). Two claudin genes found to be positively selected are *cldnd* and *cldn4*. Yet another protein family involved in tissue remodeling are the cadherins, as these cell-cell adhesion proteins are involved in transducing the mechanical tension that regulates tissue remodeling (Hinz and Gabbiani 2003; Twiss and de Rooij 2013). We found one cadherin (*pcdh10*) to be positively selected in *H. formosa*, and another cadherin (*cdh1*) to be partially duplicated in *H. formosa*. Because the duplicated *cdh1* is a modular protein, consisting of six similar cadherin domains, a partial duplication could result in a functional protein. Both of these cadherins are expressed in the mammalian placenta, with *cdh1* being essential for placental development in mice (Wolverton and Lalande 2001; Aplin et al. 2009; Stemmler and Bedzhov 2010). Finally, *btbd7* is involved in tissue remodeling as a key regulator of cleft formation in branching morphogenesis (Onodera et al. 2010). In mammals, branching morphogenesis is an important mechanism in placental development (Cross et al. 2006). As for poeciliids, much less is known about the mechanisms that regulate placenta formation, although cleft-like structures can be observed inside the folds of the follicular epithelium and branched microvilli in extensive matrotrophs (Grove and Wourms 1991; Kwan et al. 2015). GO terms associated with these genes were also significantly enriched in positively selected genes in *H. formosa*, most notably “cell-cell adhesion via plasma-membrane adhesion molecules”, and “negative regulation of endopeptidase activity” (supplementary table 2.2).

Molecular pathways other than those involved in tissue remodeling will also have played a role in placental development. For instance, a thicker follicular epithelium may result from an increased proliferation of the epithelial cells in *H. formosa*. Two of the positively selected genes identified are involved in endometrial cell proliferation in humans, namely *rbl1* and *kiaa1324* (Cavallotti et al. 2001; Deng et al. 2010).

Previous studies have shown that matrotrophic species carry an increased number of vesicles in their placental epithelial cells that are involved in trafficking nutrients from mother to embryo (Olivera-Tlahuel et al. 2018). We found one gene involved in the regulation of vesicle trafficking to be positively selected in *H. formosa*, *tmem230*. The involvement of Tmem230 in vesicle trafficking, however, has so far only been assessed in the brain (Kim et al. 2017). *Tmem230* is expressed in the human placenta (Uhlén et al. 2015), but there is no literature on the function of *tmem230* in this tissue.

These results give us a first insight into the genes that may be involved in the evolution of the placenta in *H. formosa*. Future studies should focus on generating genomic information for more species from different matrotrophic lineages in the family Poeciliidae (Reznick et al. 2002). Since the statistical power to detect positive selection is directly related to the number of species from different independent evolutionary lineages, adding genome information of more matrotrophic species and their closely related lecithotrophic 'sister-species' is likely to allow the detection of more matrotrophy-associated genes under positive selection. For example, an earlier study detected positive selection on the poeciliid *igf2* gene using the protein-coding sequence of 38 teleost species (including 26 poeciliids), of which eight are extensive matrotrophs (O'Neill et al. 2007). In our study, positive selection was not shown for *igf2* (p = 0.12). This result may be a consequence of a different role of *igf2* in *H. formosa* compared to other placental taxa, as it was shown that variation in *igf2* expression is not correlated with changes in offspring size in *H. formosa* (Schrader and Travis 2012). However, this different result may also be because using less (matrotrophic) species in the comparison reduces statistical power. In any case, genomic information for additional species will likely reveal other genes subject to positive selection that may have gone undetected in the present study. Additionally, this could also yield new insights into whether placental evolution in the different independent matrotrophic lineages is the result of selection on related or even the same genes, which would be an example of parallel evolution.

Finally, the low number of true duplications found in *H. formosa* reflects the difficulty of identifying duplicated segments using short read data only. To increase the amount of gene duplications that can be found, a reference genome of a matrotrophic poeciliid using long read or scaffolding information would be highly beneficial. Nevertheless, we were able to identify 18 genes that are exclusively selected in a highly matrotrophic species. Of these genes a high proportion is important in mammalian placenta function, suggesting convergence in the genetic building blocks of placental development between distantly related vertebrate lineages.

2.4 Conclusions

We found 18 genes that show evidence of positive selection exclusively for the branch leading to the matrotrophic species *Heterandria formosa*, and not in any of the three lecithotrophic species in the family Poeciliidae that were used for comparison. Additionally, five (partial) gene duplications were identified in *H. formosa*. A substantial portion of these genes is involved in endometrial tissue remodeling and endometrial cell proliferation, consistent with morphological changes in the placenta of *H. formosa*. Based on these results, we hypothesize that the differences in placental morphology between lecithotrophic and (extensively) matrotrophic poeciliids are at least partly due to positive selection on genes involved in tissue remodelling and endometrial cell proliferation.

2.5 Methods

2.5.1 Whole genome sequencing of *Heterandria formosa*

H. formosa individuals were caught from Wakulla Springs under state permit number 07040111, after which they were transported to Leiden, the Netherlands, where they were kept in population tanks. An F3-generation female was sacrificed using a lethal dose of ms-222. DNA was isolated from the liver using the DNeasy kit from Qiagen, according to the manufacturers' protocol. 1000 ng of DNA was sheared to a 100-800 bp range using a Covaris S-series sonicator. Genomic fragments were fit with adapters using the Paired-End DNA Sample Preparation Kit PE-102-1002 (Illumina inc.) and size-selected for 500 bp. Concentration and size profiles were determined on a Bioanalyzer 2100 using a High Sensitivity DNA chip. Paired-end sequencing was performed on an Illumina HiSeq 2000 sequencing system (Illumina Inc.) using the HiSeq Paired-End Cluster Generation Kit (PE-401-1001) and HiSeq Sequencing kit (FC-401-1001), yielding ~40X coverage of paired-end sequencing data.

2.5.2 *Heterandria formosa* genome assembly

A *de novo* assembly of the genome of *H. formosa* was made using SPAdes 3.10.0 (Bankevich et al. 2012), with default settings. To estimate the genome size and heterozygosity beforehand, we performed k-mer counting (k=20) using the Jellyfish software (Marçais and Kingsford 2011). Redundant contigs due to heterozygosity of the sample were removed using redundans v0.13c (Pryszcz and Gabaldón 2016) using default settings, and this tool was also used to rescaffold the assembly using paired-read information. After finishing of the assembly, we recalculated heterozygosity by mapping back the reads to the assembly with BWA 0.7.15 (Li and Durbin 2009), removing PCR duplicates using SAMtools 1.5 (Li et al. 2009), realigning using GATK 4.0 (McKenna et al. 2010), before variant calling using the SAMtools mpileup and bcftools call commands (Li et al. 2009), using default settings.

2.5.3 Coding sequence alignments

Published reference genomes of *Poecilia reticulata*, *Poecilia formosa* and *Xiphophorus maculatus* were downloaded from the NCBI ftp server. A scan for orthologs between these genomes was performed using ProteinOrtho 5.16 (Lechner et al. 2011), with settings -p=blastn+ and –sim=0.8. We chose to only select 1:1:1 orthologs, of which we found 8,056. In order to locate these genes in the genome of *H. formosa*, a 1:1 alignment of the *H. formosa* assembly to the *P. reticulata* genome was created using LAST 810 (Kiełbasa et al. 2011), meaning that every nucleotide from the *H. formosa* genome can align to no more than one nucleotide of the *P. reticulata* genome, and vice versa. For all selected orthologs, the *H. formosa* sequence was then retrieved via this alignment. Only genes for which the coding sequence was completely covered by the whole genome alignment were selected for further analysis, which was the case for 6,774 genes. For these genes, four-way codon alignments of the coding sequence were made using PRANK v.170427 (Löytynoja and Goldman 2008).

2.5.4 Detecting positive selection

To detect positive selection, the codeml program of the PAML (Yang 2007) package was used. This program provides a number of methods to detect positive selection, based on the ratio of non-synonymous *versus* synonymous substitutions (d_n/d_s), in the context of a known phylogenetic framework. A phylogenetic tree of the four species was constructed based on a PRANK alignment of the mitochondrial cytochrome b gene. For a neutrally evolving sequence, no distinction between synonymous and non-synonymous mutations is expected, and the d_N/d_S ratio would approach 1. Protein-coding genes however, are expected to be conserved, so purifying selection against non-synonymous mutations is expected ($d_N/d_S << 1$). Indeed, on average, protein-coding genes have a d_N/d_S ratio far below 1. However, certain situations can favor synonymous changes in a protein, for instance when a protein acquires a new (sub)function. This phenomenon is called positive selection and can lead to elevated d_N/d_S ratios at some sites in the sequence, or branches in the phylogeny. PAML provides a number of models to test for the hypothesis that a gene is evolving under positive selection. For all analyses, we deleted columns with gaps in the alignment prior to analysis by using the PAML “cleandata” function. Although this leads to somewhat conservative results, it reduces false positives due to alignment gaps.

For this study, we use two models. Firstly, we use the site model to detect genes, which contain sites subject to positive selection across the entire phylogeny. For this, we compare the fit of a model allowing $d_N/d_S > 1$ at certain codons in the coding sequence (model = 0, NSsites = 2) to a model where d_N/d_S is not allowed to go above

1 (model = 0, NSsites = 1). The assumption is that genes subject to positive selection across the whole phylogeny are not likely to be matrotrophy-associated, as three out of four investigated species are lecithotrophic. Secondly, we use the branch-site model to test for positive selection in the phylogenetic branch leading to *H. formosa*. Here, again, a model allowing $d_N/d_S > 1$ was compared to a model in which this is not the case, with d_N/d_S able to vary within both amino acid positions and phylogenetic branches (model = 2, NSsites = 2, fix_omega = 0 for the selection model, model = 2, NSsites = 2, fix_omega = 1 for the neutral model). We chose *H. formosa* as the foreground branch, testing positive selection for this phylogenetic branch only.

P-values were obtained by performing likelihood ratio tests using a chi-square distribution (df=2 for the site model, df=1 for the branch-site model, as suggested in the PAML manual). Correction for multiple testing was performed using the Benjamini-Hochberg procedure (Benjamini and Hochberg 1995), with 10% False Discovery Rate (FDR). Genes displaying significant positive selection in the branch leading to *H. formosa* were only kept in the analysis if they did not display significant positive selection for the site model. As an extra check, remaining genes were also tested for positive selection in all other branches of the phylogeny, and excluded from further analysis when this was the case. For positively selected genes belonging to gene families, 1:1 orthology was validated by aligning the *H. formosa* sequence against different *P. reticulata* paralogs of the corresponding gene family, so a false assessment of positive selection due to alignment to paralogs could be ruled out. Expression of identified genes was examined by performing blastn searches against the published transcriptome of the closely related *P. prolifica* (Jue et al. 2018), searching only against the placental and ovarian transcripts (supplementary table 2.1).

2.5.5 GO term enrichment analysis

GO terms enriched in genes subject to positive selection were detected using GOrilla (Eden et al. 2009). GOrilla takes a ranked list of genes and looks for GO terms occurring densely at the top of this list. For this, genes were ranked based on their p-value from the branch-site test, using *H. formosa* as foreground branch. We chose this method because the amount of positively selected genes in *H. formosa* was too small to find any enriched GO terms using ‘classical’ enrichment analysis (e.g. enriched GO terms in a list of significant genes compared to a background list).

2.5.6 Detecting gene duplications

H. formosa sequencing reads were mapped on the *P. reticulata* genome using BWA 0.7.15 (Li and Durbin 2009). Duplicate read removal and realignment was performed using GATK 4.0 (McKenna et al. 2010). Breakpoints in the genome indicating potential copy number variations (CNVs) were detected by running Lumpy 0.2.13 (Layer et al. 2014) on the resulting alignment file. Because of the phylogenetic distance between *H. formosa* and *P. reticulata*, regions in the genome containing no mapped reads due to sequence divergence could not be distinguished from regions deleted in *H. formosa*. As a result, deletions could not be reliably assessed. Therefore, we focused on duplicated segments. Potential duplications were validated by comparing the read depth in a potentially duplicated segment to the average read depth of the alignment file, followed by visual evaluation in JBrowse (Skinner et al. 2009b). For the validation of a potential duplication, four criteria were used. First, the read depth signal had to be at least 2 times the average read depth of the genome. Second, the coverage inside the putative duplication had to be even, that is, no coverage spikes because of repetitive elements that increase the average read depth. Third, a clear breakpoint on both sides of the CNV with discordant reads had to be visible. Fourth, only duplications with a minimum length of 1 kb were considered. To generate a set of *H. formosa*–specific duplications, resequencing libraries of *P. reticulata*, *P. formosa* and swordtail (*Xiphophorus hellerii*) were downloaded from GenBank, and the same analysis was performed for these species. If a putative duplication found in *H. formosa* was also found in one of these species, it was excluded from further analysis. Expression of genes that overlap with a duplication was examined by performing blastn searches against the published transcriptome of the closely related *P. prolifica* (Jue et al. 2018), searching only against the placental and ovarian transcripts (supplementary table 2.1).

2.6 Declarations

2.6.1 Ethics approval and consent to participate

This study was approved by the Leiden University Animal Experiments Committee under permit number 12188. *H. formosa* individuals were caught from Wakulla Springs under state permit number 07040111.

2.6.2 Consent for publications

N/A

2.6.3 Availablity of data and materials

The online version of this article (10.1186/s12862-019-1484-2) contains supplementary material, which is available to authorized users. Additionally, all supplementary material referenced to in this thesis are publicly available on the Zenodo database (10.5281/zenodo.5647272).

The sequence reads generated and analysed during the current study are available in the European Nucleotide Archive, under accession number PRJEB28818.

2.6.4 Competing interests

The authors declare that they have no competing interests.

2.6.5 Funding

HvK was supported by a WIAS research grant from Wageningen University awarded to HJM and BJAP. JvdH and BJZ were supported by European Union grants Network of Excellence LifeSpan FP6 036894 and IDEAL FP7/2007-2011/259679. In addition, JvdH was supported by a PE & RC research grant from Wageningen University. BJAP was supported by a VIDI grant (864.14.008) from the Netherlands Organisation for Scientific Research (NWO-KNAW). The funding bodies had no role in study design, analysis and interpretation, or writing the manuscript.

2.6.6 Acknowledgements

Not applicable

2.6.7 Authors' contributions

Study was designed by HJM and BJAP. JT supplied H. formosa individuals. JvdH performed sample preparation, supervised by BJZ. KK facilitated the sequencing. HvK performed the analyses in collaboration with JvdH, supervised by HJM and BJAP. HvK wrote the initial manuscript, supervised by HJM and BJAP. JT, JvdH, KK and MAMG critically reviewed the initial manuscript and provided helpful input. All authors approved the final manuscript.

3

The genomes of the livebearing fish species *Poeciliopsis retropinna* and *Poeciliopsis turrubarensis* reflect their different reproductive strategies

Henri van Kruistum[1,2], Michael W. Guernsey[3], Julie C. Baker[3], Susan L. Kloet[4], Martien A.M. Groenen[1], Bart J.A. Pollux[2] and Hendrik-Jan Megens[1,5]

[1] Animal Breeding and Genomics, Wageningen University, The Netherlands. [2]Experimental Zoology, Wageningen University, The Netherlands. [3]Department of Genetics, Stanford University School of Medicine, USA. [4]Department of Human Genetics, Leiden University Medical Center, The Netherlands. [5]Aquaculture and Fisheries, Wageningen University, The Netherlands.

Molecular Biology and Evolution (2020), 37(5), 1376-1386.

Abstract

The evolution of a placenta is predicted to be accompanied by rapid evolution of genes involved in processes that regulate mother-offspring interactions during pregnancy, such as placenta formation, embryonic development and nutrient transfer to offspring. However, these predictions have only been tested in mammalian species, where only a single instance of placenta evolution has occurred. In this light, the genus *Poeciliopsis* is a particularly interesting model for placenta evolution, because in this genus a placenta has evolved independently from the mammalian placenta. Here, we present and compare genome assemblies of two species of the livebearing fish genus *Poeciliopsis* (family Poeciliidae) that differ in their reproductive strategy: *Poeciliopsis retropinna* which has a well-developed complex placenta and *Poeciliopsis turrubarensis* which lacks a placenta. We applied different assembly strategies for each species: PacBio sequencing for *P. retropinna* (622Mbp assembly, contig N50 of 21.6 Mbp) and 10X Genomics Chromium technology for *P. turrubarensis* (597Mbp assembly, contig N50 of 4.2Mbp). Using the high contiguity of these genome assemblies and near-completeness of gene annotations to our advantage, we searched for gene duplications and performed a genome-wide scan for genes evolving under positive selection. We find rapid evolution in major parts of several molecular pathways involved in parent-offspring interaction in *P. retropinna*, both in the form of gene duplications as well as positive selection. We conclude that the evolution of the placenta in the genus *Poeciliopsis* is accompanied by rapid evolution of genes involved in similar genomic pathways as found in mammals.

3.1 Introduction

The origin of biological innovations is one of the most tantalizing questions in evolution – how does something seemingly complex emerge? All biological innovations ultimately find their origin in the genome. The evolution of genes, and the selective pressures on them to modify traits, or even generate novelties, can be studied by comparative analysis. With an increase in the number of sequenced genomes, and improvements in cost and quality, the power of comparative genomic methods is similarly increasing, bringing this approach within reach to study well established, classic, evolutionary models of biological innovation.

The fish family Poeciliidae (*sensu* (Parenti 1981)) constitutes such an evolutionary model. The Poeciliidae form a large family of live-bearing fish, consisting of approximately 275 species (Van Der Laan et al. 2014). They are widely distributed across the American continents, living in South-, Middle- and North America, as well as in the Caribbean (Reis et al. 2003). Species within this family are models for a variety of research areas, including cancer research (Meierjohann and Schartl 2006), invasion biology (Hoffberg et al. 2018), life history evolution (Reznick et al. 1996), phenotypic plasticity (Trexler et al. 1990), sexual selection (Basolo 1990) and placenta evolution (Pollux et al. 2009). Because of their use in various research areas, the genomes of several poeciliid species have been sequenced and assembled, including members of four major genera *Poecilia*, *Xiphophorus*, *Gambusia* and *Poeciliopsis* (Schartl et al. 2013; Künstner et al. 2016; Shen et al. 2016a; Hoffberg et al. 2018; Warren et al. 2018; Mateos et al. 2019).

Within the Poeciliidae, the genus *Poeciliopsis* plays a particularly important role as a model for the evolution of the placenta. Within this genus, the placenta evolved independently three times (Reznick et al. 2002). As a consequence, this genus contains closely related species that either have or lack a placenta (referred to as matrotrophic and lecithotrophic species, respectively). Additionally, the species that have placentas show a more or less continuous variation in the complexity of the placenta (Turner 1940; Kwan et al. 2015). This remarkable variation within a single genus allows for a comparison between closely related species that differ in the degree of placentation.

The implications of creating an interface between mother and offspring - that is, a placenta - have been the subject of a number of theories. For instance, it has been predicted that the evolution of the placenta should result in parent-offspring conflict: while it is in the mother's interest to balance the reproductive investments among her offspring, it is in each of the individual offspring's interest to claim more for itself than would be in the interest of the mother to give (Trivers 1974; Zeh and Zeh 2000; Crespi and Semeniuk 2004). The placenta is the site where this conflict is most apparent, because of the intimate contact between mother and offspring (Zeh and Zeh 2000). On a genomic level, parent-offspring conflict has been predicted to manifest in rapid antagonistic co-evolution of genes involved in placenta formation, embryonic development and nutrient transfer to offspring (Haig 1993; Zeh and Zeh 2000). In mammals, several studies support this hypothesis. For example, placental Cadherin (CDH) genes evolve under positive selection in humans (Summers and Crespi 2005). It is hypothesized that this rapid change is driven by antagonistic co-evolution between CDH genes and genes that modify their binding or expression, due to the influence of CDH genes on nutrient transfer from mother to offspring. In addition to positive selection, it has been shown that this rapid evolution of placental genes can also manifest in the form of gene duplications (Knox and Baker 2008). Examples of these duplications are for instance the duplication of BMP8 in mice (Zhao and Hogan 1996) and the expansion of the pregnancy-associated glycoprotein (PAG) gene family in artiodactyl species (Hughes et al. 2000). Within the Poeciliidae, evidence for this rapid genomic change as a consequence of placenta evolution is scarce (but see (O'Neill et al. 2007)).

Another predicted consequence of placenta evolution is a shift from pre-copulatory sexual selection to post-copulatory or even post-zygotic mechanisms of selection (Zeh and Zeh 2000). Indeed, it has been shown that the evolution of the placenta in the Poeciliidae correlates with phenotypic and behavioral male traits that are associated with a reduced reliance on pre-copulatory female mate choice (e.g. an absence of, or less intense, body coloration, courtship behavior and ornamental display traits) (Pollux et al. 2014). The placenta is further associated with smaller male bodies and longer genitalia, traits that facilitate sneak mating and aid in circumventing pre-copulatory female mate choice (Pollux et al. 2014). If placental species indeed rely more on post-copulatory or post-zygotic means of selection, then this should be reflected in the selective pressures on genes involved in these processes. An example of a post-copulatory process is sperm competition, which in mammals is known to drive the rapid evolution of sperm proteins (Torgerson et al. 2002). However, it is currently unknown whether increased post-copulatory selection also drives rapid evolution of sperm proteins in placental species from the Poeciliidae.

Placental evolution is thus predicted to be accompanied by rapid evolution of genes involved in placenta formation, embryonic development and nutrient transfer to offspring, as well as sperm proteins. However, whether placentation universally results in similar selection on the same pathways is unknown – the placenta in mammals evolved only once, making it very difficult to disentangle the genomic innovations and drivers behind this evolutionary novelty. To test whether similar drivers operate in the evolution of the placenta in fish, we sequenced and assembled the genomes of two species from the genus *Poeciliopsis* (family Poeciliidae): *P. retropinna* and *P. turrubarensis*. Although these two species are from the same genus, *P. retropinna* has a complex placenta, whereas *P. turrubarensis* completely lacks a placenta (Reznick et al. 2002). Comparing these genomes allows us to identify differences between these two closely related species that may stem from selection on genes associated with placentation. We employ two novel sequencing techniques to sequence these genomes: PacBio long read sequencing for *P. retropinna* (Rhoads and Au 2015), and 10X Genomics linked read sequencing, generating 'pseudo long reads', for *P. turrubarensis* (Weisenfeld et al. 2017). Both of these techniques allow for very high contiguity assemblies, either by generating very long reads (PacBio) or by using barcodes to mark sequencing reads originating from the same DNA molecule (10X Genomics).

As more high-quality genomes become available, it is becoming increasingly clear that some evolutionary novelties are based on structural variations, such as gene duplications. The high contiguity expected from our study should enable us to investigate such gene duplications and assess their potential role in generating evolutionary innovations such as the placenta. Additionally, genome annotations generated in this study allow us to perform a genome-wide scan for genes evolving under positive selection in these two genomes. Using these methods, we can test the hypothesis that the evolution of the placenta in the genus *Poeciliopsis* is accompanied by rapid evolution of genes involved in placenta formation, embryonic development, nutrient transfer to offspring and sperm proteins.

3.2 Results

3.2.1 Genome assemblies

We sequenced and assembled the genomes of *P. retropinna* and *P. turrubarensis*, using two different sequencing techniques. The genome of *P. retropinna* was sequenced and assembled using ~80X coverage PacBio sequencing, supplemented with ~40X coverage Illumina 150 bp paired end data for assembly polishing. The genome of *P. turrubarensis* was sequenced and assembled using 10X Genomics linked reads, sequenced on an Illumina platform. Both assemblies show good quality metrics. Total assembled bases are congruent with estimated genome size based on k-mer analysis, with the *P. turrubarensis* genome estimated to be slightly smaller than the *P. retropinna* genome (table 3.1). Both genomes show good contiguity, although the *P. retropinna* assembly is more contiguous than the *P. turrubarensis* assembly, having a scaffold N50 of 21.6 Mb, with half of the assembly contained within just 13 scaffolds. A blast search for telomeric repeat sequences showed that for three scaffolds, telomeres could be found on both sides, indicating that these scaffolds were completely assembled chromosomes (Figure 3.2B, Supplementary figure 3.1). In terms of expected genes, both genome assemblies are very complete: 97.5% and 95.5% of universal single-copy orthologs are present and full-length in the assembly of *P. retropinna* and *P. turrubarensis*, respectively (table 3.1). The genomes of *P. retropinna* and *P. turrubarensis* have similar repeat contents of 20.8% and 18.5%. This is comparable to other sequenced poeciliid species (Künstner et al. 2016; Hoffberg et al. 2018).

Table 3.1 Assembly statistics for the *P. retropinna* and *P. turrubarensis* assemblies.

	P. retropinna	*P. turrubarensis*
Assembly size	621.8 Mb	597.0 Mb
Predicted genome size (k-mer analysis)	621 Mb	586 Mb
Scaffold N50	21.6 Mb	4.2 Mb
Scaffold L50	13	35
Scaffold N90	6.0 Mb	0.34 Mb
Scaffold L90	31	195
Largest scaffold	30.7 Mb	17.0 Mb
Total number of scaffolds	78	5398
GC content	38.8%	39.5%
Repeat content	20.8%	18.5%
Heterozygosity	0.22%	0.34%
BUSCO score	97.5%	95.5%
Number of predicted genes	25,375	24,077

Synteny between the two assemblies is highly conserved, and large collinear blocks can be observed when aligning the assemblies (supplementary table 3.1). When aligning the five largest scaffolds from *P. turrubarensis* to the *P. retropinna* assembly, all but one of these scaffolds fall within a single *P. retropinna* scaffold, without signs of chromosomal rearrangements (figure 3.1A). The five largest *P. retropinna* scaffolds align to a number of *P. turrubarensis* scaffolds each, which is a logical consequence of the difference in contiguity between the two assemblies. Again, almost no large-scale rearrangements are evident (figure 3.1B). By contrast, when aligning the *P. retropinna* assembly to the published *Poecilia reticulata* genome assembly, many within-chromosome inversions and rearrangements can be observed. However, on the inter-chromosomal level, synteny is still highly conserved (figure 3.1C).

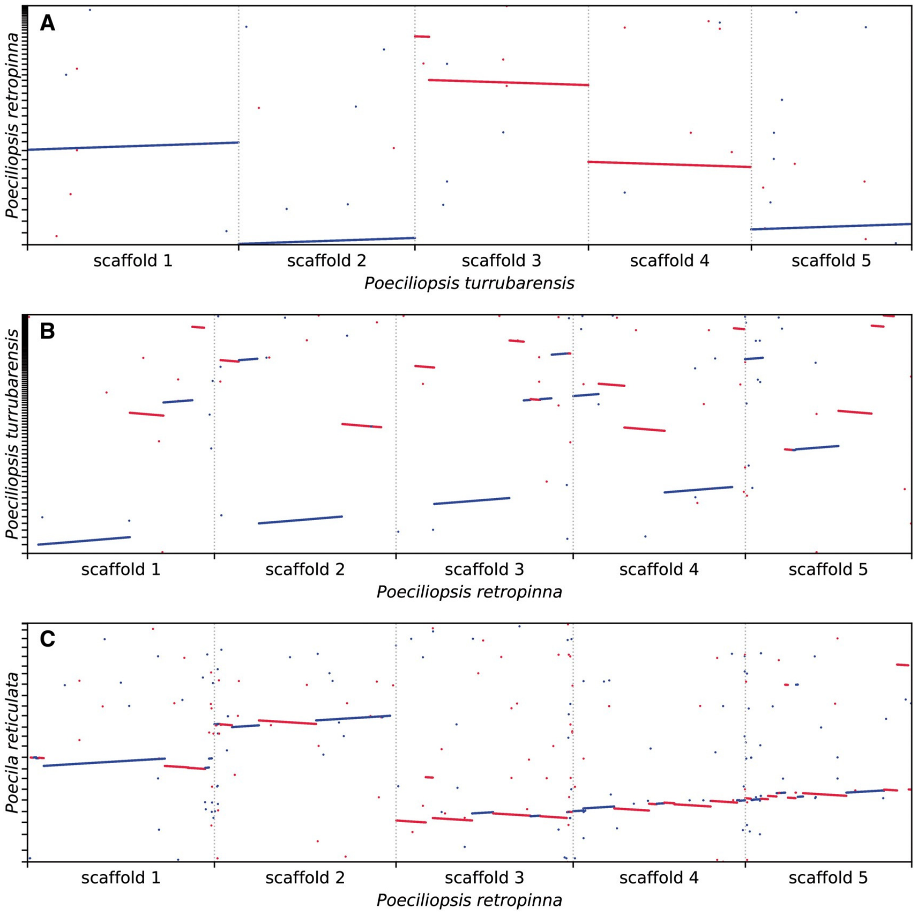

Figure 3.1 A: Dotplot of the five largest scaffolds of the *P. turrubarensis* assembly, compared to the *P. retropinna* assembly. **B**: Dotplot of the five biggest scaffolds of the *P. retropinna* assembly, compared to the *P. turrubarensis* assembly. **C**: Dotplot of the five biggest scaffolds of the *P. retropinna* assembly, compared to the published *Poecilia reticulata* assembly. Only alignments longer than 1 Kb from scaffolds longer than 1 Mb were plotted. Color codes: blue: forward alignment, red: reverse alignment.

3.2.1 Genome annotation

The genome assemblies of *P. retropinna* and *P. turrubarensis* were annotated using the BRAKER2 pipeline (Hoff et al. 2019), with subsequent filtering steps to filter out predicted genes that were likely false positives (see methods). In total, 25,375 protein-coding genes were predicted for *P. retropinna*, and 24,077 for *P. turrubarensis*. These numbers are slightly higher compared to genome assemblies from the related *Poecilia reticulata* and *Xiphophorus maculatus* (Schartl et al. 2013; Künstner et al. 2016), but similar to *Poecilia formosa* and other members from the subgenus *Mollinesia* (Warren et al. 2018). The annotations of the *P. retropinna* and *P. turrubarensis* genomes contained 94.5% and 95.4% of all universal single-copy orthologs, respectively, as determined by running BUSCO on the predicted transcriptome.

We characterized repeat content and composition using the RepeatModeler and RepeatMasker programs (Smit and Hubley 2008; Smit et al.). Furthermore, satellite DNA was identified using Tandem Repeat Finder (Benson 1999). Repeat content is similar for both genome assemblies (20.8% for *P. retropinna*, 18.5% for *P. turrubarensis*). DNA elements are the main repeat class for both genomes (Supplementary table 3.2). When looking at the repeat content throughout the largest scaffolds, some repetitive “hotspots” can be observed near the edges of these scaffolds (Figure 3.2C, Supplementary figure 3.1). Differences in occurrence are even more pronounced when looking only at satellite DNA (figure 3.2D, Supplementary figure 3.1). Regions with high density of satellite DNA likely correspond to centromeric regions, as centromeric DNA is usually satellite-rich (Lee et al. 1997). Additionally, the location of these satellite “hotspots” corresponds well to the most likely location of the centromeres, as chromosomes of Poeciliidae species are acrocentric (Cimino 1973; Haaf and Schmid 1984).

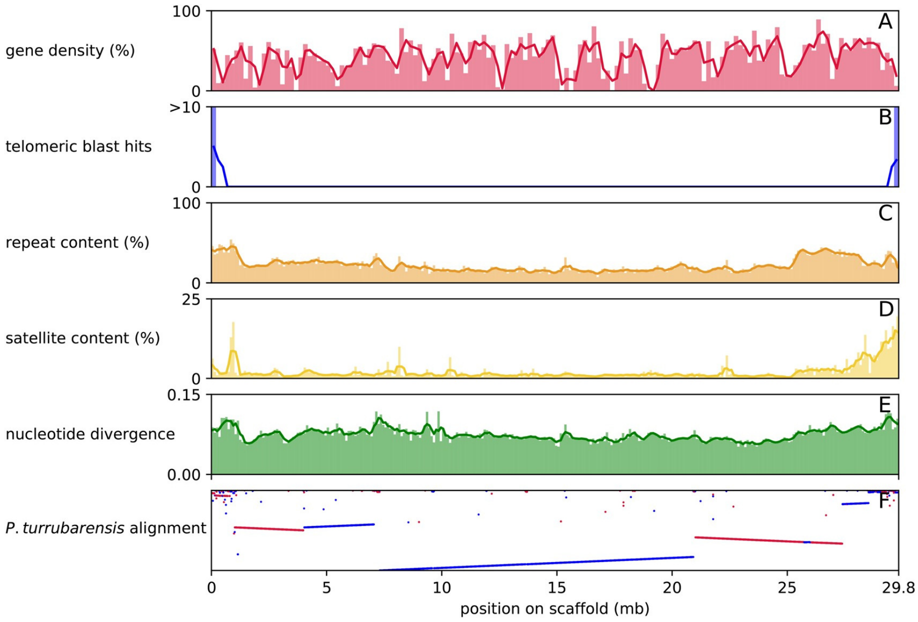

Figure 3.2 Overview of scaffold 2 of the *P. retropinna* assembly, a presumed completely assembled chromosome. Statistics are displayed in bins (bars) and moving average (line). **A**: gene density as percent of bins covered by genic region, 200kb bins. **B**: blast hits for telomeric repeats, 200kb bins. **C**: repeat content as identified by RepeatMasker, 100kb bins. **D**: satellite content as identified by Tandem Repeat Finder, 100kb bins. **E**: Pairwise nucleotide divergence between *P. retropinna* and aligned *P. turrubarensis* assembly, 100kb bins. **F**: Dotplot between *P. retropinna* scaffold 2 and aligned scaffolds of the *P. turrubarensis* assembly.

3.2.3 Whole genome alignment and gene duplications

We aligned the genome assemblies of *P. retropinna* and *P. turrubarensis* to each other and to the published genome of the guppy, *Poecilia reticulata,* (Künstner et al. 2016) using the reference free whole genome aligner ProgressiveCactus (Paten et al. 2011). The resulting alignments were post-processed into pairwise synteny blocks (see methods for details). Pairwise average nucleotide divergence between the *P. retropinna* and *P. turrubarensis* assemblies were computed, and likewise between these two assemblies and the *Poecilia reticulata* assembly (figure 3.2, supplementary figure 3.2). Nucleotide divergence between pairs of species was higher in regions with high repeat density. On average, nucleotide divergence between *Poecilia reticulata* and *P. turrubarensis* was higher than nucleotide divergence between *Poecilia reticulata* and *P. retropinna* (supplementary figure 3.2), although both species pairs have the same most recent common ancestor (Pollux et al. 2014).

Gene duplications are an important driving force in adaptive evolution. Duplicated segments from the whole genome alignment were extracted and checked for duplicated genes. Using this method, we identified 151 duplicated segments in the *P. retropinna* genome, and 102 duplicated segments on the *P. turrubarensis* genome. On these segments, 29 duplicated genes were identified in *P. retropinna*, and 4 duplicated genes in *P. turrubarensis* (supplementary table 3.3).

Within the duplicated genes in *P. retropinna*, we found a number of genes that code for sperm-associated proteins. Three duplicated proteins (spag6, cfap58 and cep162) are a part of sperm flagella or cilia in mammals (Sapiro et al. 2002; Wang et al. 2013; Nixon et al. 2018), and their expression (in humans) is either restricted to testis (*spag6* and *cfap58*), or biased towards testis (*cep162*) (Uhlén et al. 2015). One more gene that is duplicated in *P. retropinna* (*gcnt1*) is involved in sialyl Lewis X antigen presentation, which is the antigen responsible for sperm-to-egg cell recognition (Pang et al. 2011).

Another interesting result is the expansion of the Vitellogenin (*vtg*) gene family in *P. retropinna*. Vitellogenin is an egg yolk precursor, but in placental teleosts may be involved in post-fertilization nutrient provisioning (Vega-López et al. 2007). Other poeciliid fish have three *vtg* genes: two similarly sized *vtg* genes that are placed in tandem (*vtg1* and *vtg2*), and a smaller *vtg* on a different location on the same chromosome (*vtg3*). However, in *P. retropinna*, *vtg1* is duplicated, leading to a cluster of three *vtg* genes placed in tandem, along with the additional *vtg3* gene (figure 3.3).

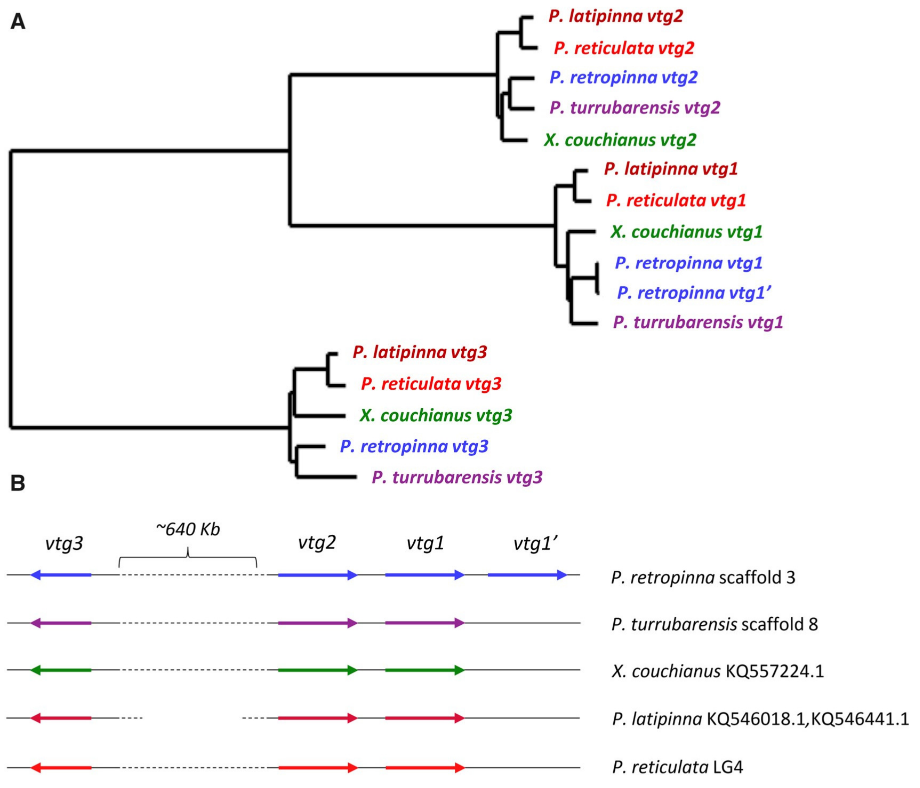

Figure 3.3 A: Phylogeny of the *vtg* gene family in five poeciliid fish species. **B**: conserved synteny of the *vtg* gene cluster in five poeciliid species.

3.2.4 Positive selection

We performed likelihood tests for positive selection in phylogenetic branches leading to *P. retropinna* and *P. turrubarensis* using the codeml program from the PAML package (Yang 2007). Briefly, the hypothesis was tested that a coding sequence shows an elevated ratio of non-synonymous to synonymous mutations at certain codon sites, in one of the tested branches, implemented in PAML as the branch-site model. For this, we used 1:1 orthologous genes found in four poeciliid genomes, of which the coding sequence was aligned using a codon-aware aligner (see methods). We found 180 genes evolving under positive selection in *P. retropinna*, and 165 genes evolving under positive selection in *P. turrubarensis*, using a 5% false discovery rate (supplementary table 3.4). GO term enrichment analysis showed that the genes evolving under positive selection in *P. retropinna* were significantly enriched for genes involved in “neural system development”, and "protein targeting to peroxisome" (supplementary table 3.5). We found that five out of seven genes with GO term "protein targeting to peroxisome" in this gene set are involved in peroxisomal lipid metabolism, while the other two genes are involved in regulating the transport of enzymes to the peroxisomes.

Shown in figure 3.4 is the probability of positive selection for each amino acid of one of these peroxisomal enzymes, peroxisomal multifunctional enzyme type 2 (*hsd17b4*). This enzyme is involved in peroxisomal beta oxidation of fatty acids. Positive selection is most apparent in three subsequent residues inside the N-terminal MaoC-like dehydratase domain. This domain contains no catalytic sites, but instead determines substrate specificity (Koski et al. 2004). Therefore, positive selection within this domain may change the substrates on which the enzyme acts, without impairing catalytic function.

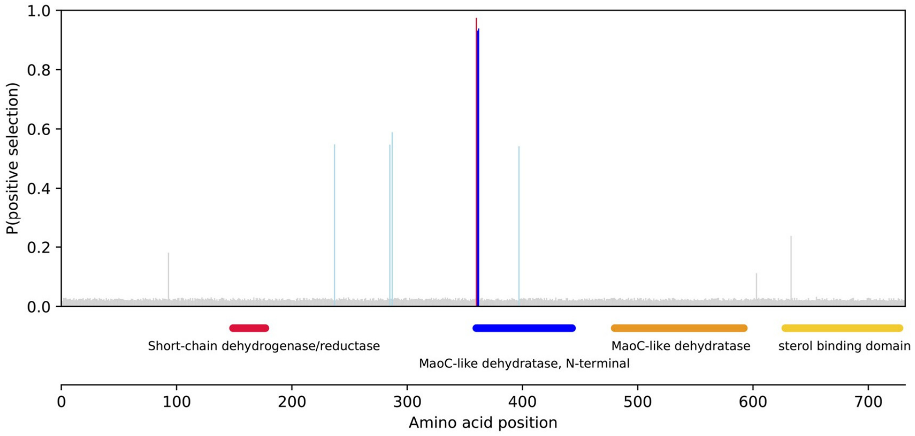

Figure 3.4 Likelihood of positive selection for each codon in *P. retropinna* peroxisomal multifunctional enzyme 2. Conserved domains are plotted on the bottom panel. Color codes for probability of positive selection: Red > 95% > blue > 90% > light blue > 50% > grey.

Additionally, we found several genes coding for parts of the Wnt signaling pathway to be positively selected in *P. retropinna* (*wnt6*, *wnt7b*, *cxxc4*, *rspo2*, *ctnnd1*), suggesting changes in this pathway as well. The Wnt signaling pathway is an essential signaling cascade in embryonic development throughout the vertebrate lineage (Clevers 2006), and a regulator of placenta formation in mammals (Sonderegger et al. 2010).

For *P. turrubarensis*, significant enrichment among positively selected genes was found for the GO term "phosphate-containing compound metabolic process" (supplementary table 3.6). Genes belonging to this GO term seemed to be involved in a variety of metabolic processes and pathways, in which we did not observe a clear trend.

It is predicted that mate choice prior to copulation plays a more important role in sexual selection in non-placental (lecithotrophic) species compared to their placental relatives. In our analysis, we find four genes involved in vision that are positively selected in *P. turrubarensis* (*eya4*, *rlbp1*, *opso*, *rp9*). Positive selection on genes involved in vision could facilitate optimizing signaling of visual cues from potential mating candidates. Additionally, we found evidence of positive selection in *P. turrubarensis* for the Melanin Concentrating Hormone Receptor 1 (*mchr1*) gene. Melanin concentrating hormone is involved in pigment pattern formation in teleosts (Nagai et al. 1986).

3.3 Discussion

Studying the origin of biological innovations by comparative approach requires high-quality genomes and fairly complete gene sets. To meet this requirement, we sequenced and assembled the genomes of two live-bearing fish from the same genus: *Poeciliopsis retropinna* and *Poeciliopsis turrubarensis*. The resulting genomes are of high contiguity and quality, and will be valuable for future genomic analyses that involve these species.

Although both assemblies showed good quality metrics, the *P. retropinna* assembly, based on PacBio sequencing, outperformed the *P. turrubarensis* assembly both in terms of contiguity and gene completeness, especially in regions that are rich in repeats. (figure 3.2, supplementary figure 3.1). An additional advantage of the PacBio assembly is that we could validate tandem duplications directly by mapping reads to the assembly (see supplementary figure 3.3 for example). However, these advantages come at a cost, as PacBio is considerably more expensive in raw sequence generation and computational resources for assembly, even with an efficient assembler such as wtdbg2. A possible reason for the difference in contiguity is that 10X Genomics based assemblies, even more so than PacBio based assemblies, rely heavily on the extraction of very high molecular weight DNA. DNA extraction based on red blood cells allows for the highest molecular weight DNA, since no mechanical degradation of tissue is necessary for DNA extraction. Sampling blood, however, is not feasible for these small fishes as single fish simply do not contain sufficient quantities. Extracting DNA from tissues, as was done in this study, results in additional mechanical stress on DNA, resulting in a shorter average molecule size (~40kb). Finally, the 10X Genomics method is primarily optimized for mammalian genomes; non-mammalian genomes are known to result in reduced contiguity. Despite the observed difference in contiguity, the quality of both genomes allows for a comprehensive and systematic assessment of gene duplications, as well as a genome-wide scan for genes evolving under positive selection.

The evolution of the placenta in mammals was accompanied by an increased rate of evolution of genes involved in mother-offspring interaction: genes involved in nutrient transfer to offspring, embryonic development and placenta formation (Clark et al. 2003; Crespi and Semeniuk 2004; Nielsen et al. 2005). Furthermore, sperm selection is predicted to intensify in placental species (Zeh and Zeh 2000) and is hypothesized to drive rapid evolution of sperm proteins in placental mammals (Torgerson et al. 2002). Our results indicate that similar processes are ongoing in the placental *P. retropinna*.

The high rate of evolution of genes involved in nutrient transfer in *P. retropinna* may provide insights into the metabolic mechanisms that play a role in the evolution of the poecilid placenta. Specifically, five enzymes involved in peroxisomal lipid metabolism evolve under positive selection in *P. retropinna*, as well as two genes involved in protein transport to the peroxisomes, highlighting a very specific metabolic pathway. Lipids metabolized in the peroxisome include very long chain fatty acids and certain steroids (Mannaerts and Van Veldhoven 1996). The expression of these enzymes is regulated by the Peroxisome Proliferator Receptor α (PPARα) nuclear receptor (Reddy and Chu 1996), and it has been reported that PPARα regulates placental lipid metabolism through the proliferation of peroxisomes in rats (Martinez et al. 2008). Co-upregulation of peroxisomal lipid metabolizing enzymes and Vitellogenin was observed after treatment with estrogenic compounds in Zebrafish (Ortiz-Zarragoitia and Cajaraville 2005). These results may suggest that changes in peroxisomal lipid metabolism in *P. retropinna* result from a shift from pre- to post-fertilization provisioning that is inherent to the evolution of a placenta. Directly related to this may be the expansion of the Vitellogenin (*vtg*) gene family in *P. retropinna*. Vitellogenin proteins are egg yolk precursors that, in non-placental poeciliids, are synthesized in the liver before being transferred to the oocyte to build up nutrient supply, a process called vitellogenesis (Rocha et al. 2008). Placental poeciliids such as *P. retropinna* supply only a very small amount of yolk before fertilization, which makes an expansion of the *vtg* gene family seem somewhat counterintuitive at first. However, it has been shown that Vitellogenin is also used as a means of post-fertilization maternal nutrient provisioning in placental fish from the family Goodeidae (Vega-López et al. 2007). This duplication suggests that Vitellogenin may fulfil a similar function in *P. retropinna*. At the same time, the egg, although small, does contain some yolk. One of the *vtg* gene copies may be involved new role in embryo provisioning, while the other remains active in yolking the egg. Although we have no further evidence to support this, it can be hypothesized that the two copies therefore have different timing in expression.

Innovations in tissue development, as clearly the case during placental evolution, require novel recruitment of signalling- and other pathways involved in cell differentiation and cell migration. In this study we report the rapid evolution of genes involved in embryonic development in *P. retropinna*. The most striking finding are five genes involved in the Wnt signalling pathway that were found to evolve under positive selection. This pathway is crucial in early embryonic development and tissue formation, and its function is conserved across vertebrates as well as some invertebrates (Clevers 2006). By contrast, no members of the Wnt pathway were found to evolve under positive selection in *P. turrubarensis*, implying that these changes are not characteristic for the genus *Poeciliopsis* but rather a feature of the lineage leading to the placental *P. retropinna*.

Another compelling finding is that four genes encoding sperm-associated proteins are duplicated in *P. retropinna*. Three out of four of these proteins are located in sperm flagella of cilia. This may be an indication for increased post-copulatory selection in *P. retropinna*, for instance on sperm quantity or quality. However, we also find some evidence for rapid evolution of sperm proteins in the non-placental *P. turrubarensis* in the form of positive selection on three sperm-associated genes (*spata4*, *spata17*, *spag16*). These proteins are involved in sperm production (Liu et al. 2004; Nie et al. 2013) and motility (Zhang et al. 2007). Although it is predicted that sperm selection intensifies in placental species, sperm selection plays a role in non-placental poeciliids as well (Brown et al. 2018). These results indicate a continuing selection on sperm-related traits, with possible different traits related to the different life histories in the two species under study.

The rapid evolution of genes involved in mother-offspring interactions in mammals show a high degree of overlap, in the pathways involved, with accelerated evolution in genes in *P. retropinna*. On the gene level, we do not find convergence between our results and studies on the genomic consequences of mammalian placenta evolution (Clark et al. 2003; Crespi and Semeniuk 2004; Nielsen et al. 2005; Hou et al. 2009). This may be explained by the fact that while the mammalian and poeciliid placentas perform a similar function (e.g. post-fertilization nutrient transfer to offspring), the morphological and physiological characteristics of these placentas are in fact quite different (Griffith and Wagner 2017). For example, the *vtg1* gene that is duplicated in *P. retropinna*, and which may be involved in lipid transfer to offspring in that species, is not functional in mammals. In mammals lipid transfer to offspring is mediated by other proteins such as apolipoprotein B-100 (Madsen et al. 2004). Similarly, many aspects of placentation may be functionally convergent but not molecularly homologous between the poeciliid and mammalian placenta.

Genomes of non-model species are increasingly accessible for accurate characterization and annotation, creating opportunities for comparative analyses by filling in crucial phylogenetic gaps providing powerful evolutionary contrasts. This study shows compelling evidence for selection on genes and pathways involved in a key life history trait, providing a valuable hypothesis on placental evolution in livebearing fish that can be readily tested in future studies.

3.4 Methods

3.4.1 DNA preparation and sequencing

A male individual of *P. retropinna* and a male individual of *P. turrubarensis* were caught at Rio Cañas, Costa Rica. DNA was extracted using the Qiagen Genomic-tip 100/G DNA extraction kit, following the instructions on the manufacturer's protocol. For the *P. turrubarensis* sample, extracted DNA was size-selected for fragments > 30kb using a BluePippin system. The *P. retropinna* DNA was sequenced on a PacBio Sequel System (Rhoads and Au 2015), generating ~6.75 million reads with an N50 read length of ~17kb. Additionally, a fraction of the same DNA was sequenced on an Illumina HiSeq 4000 sequencer, generating ~240 million 150bp paired-end reads. The *P. turrubarensis* DNA was processed by a Chromium controller chip, together with 10x Chromium reagents and gel beads following the manufacturer's protocol. Subsequently, the DNA was sequenced in a single lane of an Illumina HiSeq X Ten sequencer, producing ~344 million 150bp paired-end reads.

For generating RNA-seq data, pregnant females were stored in RNAlater. Dissection of individual maternal follicles was performed in the laboratory and stored in TriZol. RNA was extracted from late stage maternal follicles in six *P. retropinna* and one *P. turrubarensis* females using a Direct-zol RNA miniprep kit. Total RNA was then depleted of rRNA as described by (Adiconis et al. 2013). This depletion was only partial, because human DNA oligos were used, as *Poeciliopsis*-specific DNA oligos tiling rRNAs do not exist. Libraries were subsequently prepared using the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina for use with rRNA depleted RNA. In brief, 100ng of each RNA sample was subjected to a 7-minute fragmentation and primed using random primers, followed immediately by cDNA synthesis. Samples were end-prepped and standard NEBNext adapters were ligated to the resulting cDNAs. Each sample was enriched for adapter ligated DNA using a 7-cycle PCR with each sample using a unique index primer, and then purified using Ampure XP beads. Library quality was assessed using both Qubit and Bioanalyzer technology. Libraries were multiplexed, and then sequenced on an Illumina NextSeq using one 150 cycle High Output v2 kit.

3.4.2 *Poeciliopsis retropinna* assembly

Before assembly of the *Poeciliopsis retropinna* genome, genome size and heterozygosity were estimated using a k-mer analysis. K-mers from Illumina short reads were counted using KMC 3 (Kokot et al. 2017). The assembly was generated using a hybrid pipeline where reads were corrected using Canu v1.7.1 (Koren et al. 2017) using settings genomeSize=625m, MhapSensitivity=normal, correctedErrorRate=0.085, corMhapSensitivity=normal, corOutCoverage=50, and assembly was performed by wtdbg v2.1 (Ruan and Li 2019) using default settings. After assembly, redundant contigs corresponding to heterozygous regions in the genome were removed using Redundans v0.14a (Pryszcz and Gabaldón 2016), only using the reduction step of the pipeline. Then, the assembly was re-scaffolded using SSPACE-longread v1-1 (Boetzer and Pirovano 2014). The reasons for re-scaffolding an assembly already based on long reads are twofold: first, contigs corresponding to heterozygous regions of the genome were removed by running Redundans. These sections of the genome can decrease contiguity of an assembly by breaking contigs at heterozygous parts. After these contigs were removed this was no longer an issue, allowing further scaffolding of the assembly. Second, assembly was performed after read correction by Canu, which uses all reads to correct the longest reads corresponding to 50X coverage (as determined by the corOutCoverage parameter). Other reads that are not used for the assembly may still bridge multiple contigs. After re-scaffolding, the draft assembly was polished by aligning Illumina reads to the genome using bwa v0.7.5, using the 'mem' option (Li and Durbin 2009), after which the assembly was polished in two rounds using Pilon v1.22 (Walker et al. 2014). Gene completeness was assessed using BUSCO 3.0.2 (Simão et al. 2015). The assembly was checked for possible mis-assemblies by alignment to the published guppy (*Poecilia reticulata*) genome assembly (v1.0) (Künstner et al. 2016) using MUMmer 3.23 (Kurtz et al. 2004). One putative mis-assembly was manually corrected.

3.4.3 *Poeciliopsis turrubarensis* assembly

For the *Poeciliopsis turrubarensis* assembly, Illumina basecall files were demultiplexed using bcl2fastq v2.20 (https://support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software.html), after which they were assembled using the Supernova assembler v2.1.1 (Weisenfeld et al. 2017), using setting --maxreads=300,000,000, producing an assembly with a size of 618 Mb, and a scaffold N50 of 4.2 Mb. Although the Supernova assembler should separate haplotypes for the majority of the genome, we still observed a number of redundant contigs belonging to heterozygous regions in the genome that were assembled separately. We removed these using Redundans v0.14a (Pryszcz and Gabaldón 2016), using only the reduction step of the pipeline. Gene completeness was assessed using BUSCO 3.0.2 (Simão et al. 2015).

3.4.4 Genome annotation

Placental RNA-seq libraries of *P. retropinna* and RNA-seq libraries of the corresponding follicular tissue in *P. turrubarensis* were aligned to their respective genomes using HISAT2 2.1.0 (Sirén et al. 2014). Additionally, published RNA-seq libaries of *Poeciliopsis prolifica* (Jue et al. 2018) tissues were downloaded from GenBank and aligned in the same way. For both genomes, a *de novo* repeat library was constructed using RepeatModeler 1.0.11 (Smit and Hubley 2008). Subsequently, the genome was softmasked using RepeatMasker 4.0.7 (Smit et al. 2013), using these newly created species-specific repeat libraries. The softmasked genomes were subsequently annotated using the BRAKER2 pipeline (Hoff et al. 2019). Besides the aligned RNA-seq, protein sets from *Poecilia reticulata* and *Xiphophorus maculatus* were included as additional evidence for the annotation pipeline. After annotation, gene predictions were filtered by scanning all predicted proteins for functional domains against the protein family (pfam) database (Bateman et al. 2004), using HMMer 3.2.1 (Mistry et al. 2013). Predicted genes coding for a protein without any known protein domain were excluded from the final gene set. Additionally, predicted genes having a total length shorter than 500 base pairs were also excluded from the gene set. Functional annotations were added by performing protein Blast (Altschul et al. 1990) searches to the Swissprot database (Boeckmann et al. 2003) and running InterProScan (Jones et al. 2014) to add domain information for every predicted protein.

3.4.5 Whole genome alignment

Soft-masked assemblies of *P. retropinna* and *P. turrubarensis* were aligned to each other and to the published genome assembly of *Poecilia reticulata* (Künstner et al. 2016) using ProgressiveCactus (Paten et al. 2011). Pairwise alignments were obtained by generating a Multi Alignment Format (MAF) file for pairs of species using the hal2maf utility that is part of the hal toolbox (Hickey et al. 2013). Then, synteny blocks were formed by chaining the resulting pairwise alignments using UCSCs axtChain program (Kent et al. 2003). Because the ProgressiveCactus alignment is reference-free, we could obtain these synteny blocks relative to each of the three genomes that we used, as well as from within-genome alignments (paralogous blocks). Chains were converted to GFF3 format for further analysis and visualization in Jbrowse (Skinner et al. 2009a).

3.4.6 Gene duplications

Gene duplications between the genomes of *P. retropinna* and *P. turrubarensis* were assessed using synteny blocks, applying the following criteria. First, regions of the reference genome where multiple synteny blocks of the query genome align to the same region of the reference genome were extracted. Then, these regions were filtered on the following criteria: (1) the overlapping synteny blocks should also overlap in the positions to which their “raw” alignments align, (2) the duplicated region must be at least 3 kb long, (3) no paralogous alignments should be present within the duplicated region, (4) the duplication must lie on a scaffold that is bigger than 100kb, to prevent separate assembly of heterozygous regions of the genome to mimic a duplication, and (5) the average coverage of mapped short reads within the duplicated region must not differ more than 25% from both the surrounding region (10kb upstream and downstream) and the genome-wide average. Genes that were found to reside within these duplicated regions were extracted from the genome annotation. To infer the most likely ancestral copy number of these genes, the *Poecilia reticulata* genome was included in the alignment as an outgroup. Using the same methodology as previously stated, the gene copy number was determined for *Poecilia reticulata*. Duplications were kept only if the *Poecilia reticulata* copy number matched the non-duplicated state. For an overview of the effect of distinct filtering steps on amount of duplicated segments passing the filtering, see supplementary table 3.7.

3.4.7 Positive selection

We detected orthologous genes shared between the genomes of *P. retropinna*, *P. turrubarensis*, *Poecilia reticulata* (Künstner et al. 2016) and *Xiphophorus maculatus* (Schartl et al. 2013) using ProteinOrtho v5.16b (Lechner et al. 2011). To avoid comparing paralogs, we only selected genes displaying 1:1 orthology for all species. In total, we found 11323 1:1 orthologs. A codon-aware alignment was made for the sequences of these orthologs using PRANK v.170427 (Löytynoja 2014). These alignments were tested for signs positive selection using the codeml program that is part of PAML 4.9 (Yang 2007). Alignment columns that have gaps in one of the species were excluded from analysis. All genes were tested for positive selection in the phylogenetic branch leading to *P. retropinna*, as well as in the branch leading to *P. turrubarensis* by applying the branch-site model. Additionally, we tested the genes for positive selection across the whole phylogeny using the site model. This way, a distinction was made between positive selection that is unique for a certain branch and positive selection that is shared between all poeciliid species. P-values were obtained by performing likelihood ratio tests using a chi-square distribution, following recommendations from the PAML manual. Correction for multiple testing was performed with the Benjamini-Hochberg procedure (Benjamini and Hochberg 1995), using a 5% false discovery rate. For genes with sufficient evidence for positive selection, the sites most likely evolving under positive selection were selected using the Bayes Empirical Bayes (BEB) analysis integrated in the PAML toolkit.

3.5 Additional files

The online version of this article (10.1093/molbev/msaa011) contains supplementary material, which is available to authorized users. Additionally, all supplementary material referenced to in this thesis are available on the Zenodo database (10.5281/zenodo.5647272).

Sequencing reads used for both assemblies are available in the European Nucleotide Archive under accession number PRJEB3354. Genome assemblies are available in GenBank under accession numbers PRJNA555005 for *P. retropinna* and PRJNA555006 for *P. turrubarensis*.

3.6 Acknowledgements

This work was supported by a WIAS research grant from Wageningen Institute of Animal Sciences (to HJM & BJAP), a VIDI grant from the Netherlands Organisation for Scientific Research (864.14.008 to BJAP), a National Science Foundation Graduate Research Fellowship (2014185511 to MWG) and a Diversifying Academia, Recruiting Excellence (DARE) Fellowship (to MWG).

3.7 Competing interests

The authors declare that they have no competing interests.

4

Parallel genomic changes drive repeated evolution of placentas in livebearing fish

Henri van Kruistum, Reindert Nijland, David N. Reznick, Martien A.M. Groenen,
Hendrik-Jan Megens and Bart J.A. Pollux

[1] Animal Breeding and Genomics group, Wageningen University, The Netherlands. [2]Experimental Zoology group, Wageningen University, The Netherlands. [3]Marine Animal Ecology group, Wageningen University, The Netherlands. [4]Department of Biology, University of California, USA. [5]Aquaculture and Fisheries group, Wageningen University, The Netherlands.

Molecular Biology and Evolution (2021) 38(6): 2627–2638,

Abstract

The evolutionary origin of complex organs challenges empirical study because most organs evolved hundreds of millions of years ago. The placenta of live-bearing fish in the family Poeciliidae represents a unique opportunity to study the evolutionary origin of complex organs, because in this family a placenta evolved at least nine times independently. It is currently unknown whether this repeated evolution is accompanied by similar, repeated, genomic changes in placental species. Here we compare whole genomes of 26 poeciliid species representing six out of nine independent origins of placentation. Evolutionary rate analysis revealed that the evolution of the placenta coincides with convergent shifts in the evolutionary rate of 78 protein-coding genes, mainly observed in transporter- and vesicle-located genes. Furthermore, differences in sequence conservation showed that placental evolution coincided with similar changes in 76 non-coding regulatory elements, occurring primarily around genes that regulate development. The unexpected high occurrence of GATA simple repeats in the regulatory elements suggests an important function for GATA repeats in developmental gene regulation. The distinction in molecular evolution observed, with protein-coding parallel changes more often found in metabolic and structural pathways, compared to regulatory change more frequently found in developmental pathways, offers a compelling model for complex trait evolution in general: changing the regulation of otherwise highly conserved developmental genes may allow for the evolution of complex traits.

4.1 Introduction

The emergence of complex organs is one of the most significant phenomena in the evolution of multi-cellular organisms. Characterizing the origin of this complexity is a challenge because we are most often confronted with the end-products of evolution as they appear in currently living organisms, with little knowledge on intermediate stages. Ultimately, the development of these organs is encoded in the genome. This same genome, however, poses a puzzle: while there is remarkable diversity in vertebrate morphology, the genes that regulate morphological development tend to be highly conserved (Gaunt 2002; Hoegg and Meyer 2005).

Over the past decades, developments in genome science have unraveled details in how cell differentiation, cell signaling, and cell migration shape organisms and their organs during ontogeny. As the developmental pathways leading to specific organismal traits are better understood, it is becoming clear that organs, especially in vertebrates, are not only highly conserved in morphology and physiology, but also in developmental pathways (Farrell et al. 2018). The deep conservation in developmental pathways deployed once organs have emerged in evolution, however, does question the genomic basis of convergence: if structures, such as organs, develop in parallel in another animal group, is that mirrored in convergence in underlying molecular pathways, or is the parallel evolution only superficial, based on different developmental triggers?

Studies of genomic changes associated with convergent phenotypic evolution have identified genomic changes in physiological and structural genes common to convergent lineages (Foote et al. 2015; Chikina et al. 2016). However, this alignment of convergent morphology with convergent changes in the genome does not include the developmental genes that govern morphology. The absence of the expected association between developmental genes and phenotypic evolution may be because evolution has been assessed via changes in amino acid sequences or copy number, while another cause for morphological evolution could lie in changes in the spatio-temporal expression patterns of developmental gene expression (Carroll 2008; Levin et al. 2016). Such changes in expression are instead controlled by the elements that regulate developmental genes.

An excellent model to study convergent evolution of a complex trait is found in the live-bearing fish family Poeciliidae. The Poeciliidae are a family of livebearing fish consisting of around 275 species (Parenti 1981; Van Der Laan et al. 2014). In this family, a placenta has evolved independently at least nine times from a non-placental ancestor (Pollux et al. 2009; Furness et al. 2019). The evolution of the placenta in this family coincides with a shift from pre- to post-fertilization nutrient provisioning to the offspring. Non-placental or lecithotrophic species supply nutrients to their offspring before fertilization in form of egg yolk. Placental or matrotrophic species supply only a very small amount of yolk, with the majority of nutrients being supplied after fertilization by means of their placenta. Furthermore, species that have a placenta of intermediate complexity are also present in this family, where nutrients are provided both before and after fertilization (Jollie and Jollie 1964; Grove and Wourms 1994; Kwan et al. 2015).

On a morphological level, differences between placental and non-placental poeciliid species are found in the follicular tissue surrounding the embryos. In non-placental species, a thin follicle can be observed surrounding the embryos. In placental species however, this tissue is thicker and has extensive folds (Jollie and Jollie 1964; Grove and Wourms 1991; Grove and Wourms 1994; Kwan et al. 2015). Additionally, microvilli and vesicles can be observed in the follicular tissue of placental species. The variation in placental complexity can be characterized by quantifying the degree of post-fertilization nutrient provisioning to offspring with the Matrotrophy Index (MI) (Reznick et al. 2002; Pollux et al. 2009). The MI is defined as the embryo mass at birth divided by the egg mass at fertilization, and is used as a proxy for placental complexity (Pollux et al. 2009). Morphological studies have shown that the MI correlates well with the placental complexity in multiple poeciliid species (Jollie and Jollie 1964; Grove and Wourms 1994; Kwan et al. 2015).

The placenta in the fish family Poeciliidae allows for a genomic study of complex trait evolution because of several reasons. First, the integration of estimates of MI with a DNA-based phylogeny for the family suggests that the degree of placentation can evolve very quickly, with estimations based on molecular data suggesting that a placenta can evolve in as little as 0.75 million years (Reznick et al. 2002). Placenta evolution in the Poeciliidae is a fairly recent event, with the most recent placenta evolutions in the genus *Poeciliopsis* being estimated to have happened less than five million years ago (Reznick et al. 2017). Second, in some cases, species with placentas have closely related sister species without a placenta, which allows for comparisons between species that differ in placentation (Reznick et al. 2002; Pollux et al. 2014), but have very similar genomes in general (van Kruistum et al. 2020). Third, the multiple independent evolutions of the poeciliid placenta allows investigating genomic changes between multiple instances of placenta evolution that happened in parallel, which will increase the reliability of our results.

Previous studies have identified species-specific genomic changes in pathways involved in metabolism and development in individual placental poeciliids (O'Neill et al. 2007; Jue et al. 2018; Van Kruistum et al. 2019; van Kruistum et al. 2020). However, these studies did not compare multiple instances of placenta evolution on a genome-wide scale, include closely related non-placental species, or consider non-coding regions of the genome. The similar physiological and morphological details of placentation, and the similarities in intermediary paths towards placentation, suggest that throughout the Poeciliidae similar pathways are at the basis of conferring placentation in all these species. Moreover, the intermediate stages suggest a quantitative nature of the trait, meaning it is not expected that a single 'switch' gene exists that regulates this trait. Here we take advantage of the multiple independent placenta evolutions in our study system to test whether the convergent morphological evolution of the placenta is reflected in convergent molecular evolution on a genomic level.

In this study we use both publicly available and new genome assemblies to construct a large-scale comparative framework of genome evolution in the Poeciliidae, consisting of 26 species, of which eight species have a placenta (figure 1). These eight placental species include six independent origins of placentation. Although it may seem from our data that independent losses of placentation are also a plausible option in the genus *Poeciliopsis*, studies that include more species from this genus show that the placental species we include do in fact represent three independent instances of placenta evolution (Reznick et al. 2002; Reznick et al. 2017). By comparing the genomes of these 26 species, we are able to investigate (i) which genomic changes are associated with placenta evolution in the Poeciliidae, (ii) whether similar genomic changes occur in each of the six origins of placentation, and (iii) whether mutation in coding or non-coding genomic regions are most important for placenta evolution. Our study will provide new insights in genome evolution during the evolution of complex traits.

4.2 Results

4.2.1 Reconstructing a molecular phylogeny

We reconstructed a Maximum Likelihood phylogeny of the Poeciliidae using a concatenated alignment consisting of (i) whole mitochondrial genomes acquired from the whole genome sequencing data and (ii) the complete coding sequence from 1010 nuclear genes (figure 4.1). Because of the difference in nuclear divergence between mitochondrial an nuclear DNA, these parts of the data were partitioned separately (supplementary figure 4.1). For the same reason, the three codon positions were also partitioned separately for both nuclear and mitochondrial genes. The resulting phylogeny was used as the basis for subsequent comparative analyses.

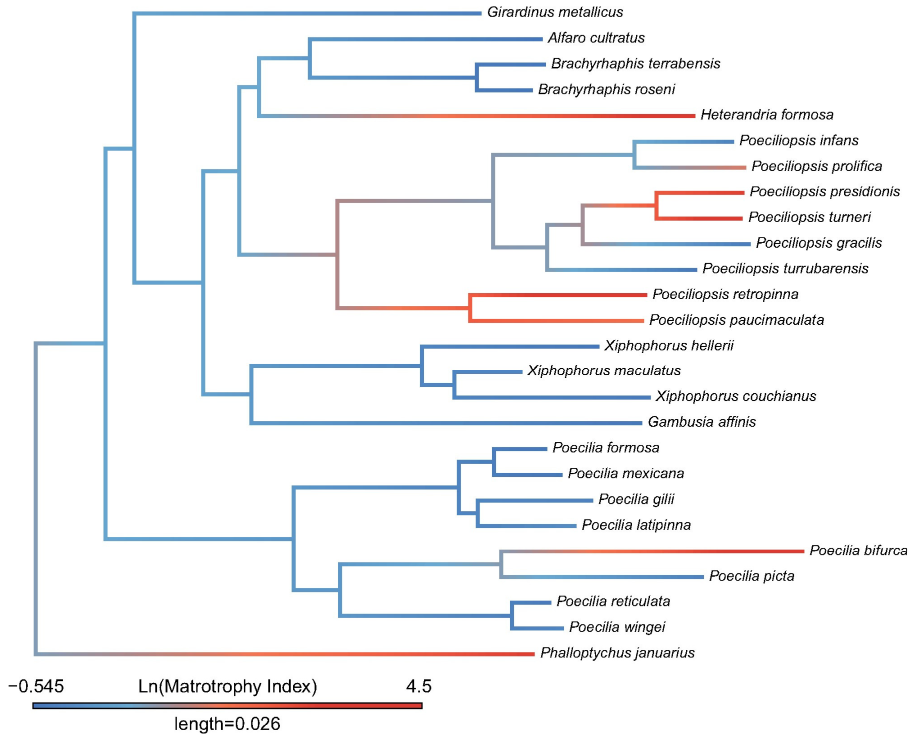

Figure 4.1 Molecular phylogeny of investigated species. The colors of the branches represent the natural logarithm of the Matrotrophy Index (MI), either the observed MI for terminal branches or the estimated MI as determined by ancestral trait reconstruction.

4.2.2 Convergent shifts in evolutionary rate for genes in placental poeciliid species

To test for the presence of convergent genomic changes in protein-coding regions, we applied an evolutionary rate analysis based on the method used by Chikina *et al*

(Chikina et al. 2016), modified to test for correlation between evolutionary rate and the continuous variable MI, instead of testing for a significant difference between two discrete classes. This analysis tests whether the relative rate of gene evolution is correlated with the MI across the phylogeny. The relative evolutionary rate is inferred from the number of amino acid substitutions across each branch of the phylogeny for a certain protein-coding gene and is normalized for branch- as well as gene-specific evolutionary rates (see methods). These relative rates are then tested for correlation with observed or estimated MI values using Spearman's correlation test. Both the observed and estimated MI values can be found in supplementary table 4.1. We created a null distribution by generating simulated data sets in which the MI values were randomly assigned to individual branches (orange in figure 4.2). We infer genome-wide convergence in the rate of evolution by comparing the distribution of p-values for the null and observed distributions. Analysis of the simulated data revealed a uniform p-value distribution, as expected, when no convergent evolution is present. However, the results for placental branches were heavily skewed towards lower p-values (figure 4.2), indicating that more genes show a good correlation between MI and evolutionary rate than would be expected by chance (p=1.36e-61, Kolmogorov–Smirnov test).

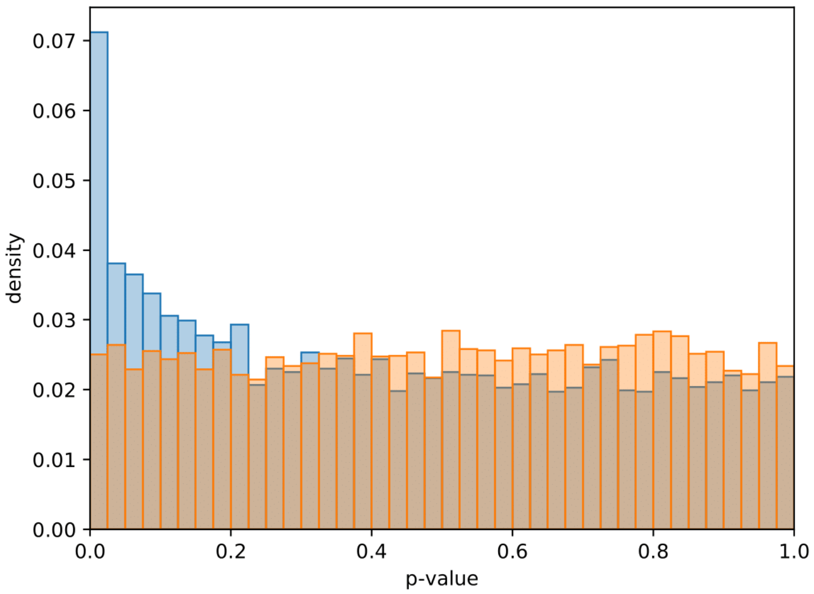

Figure 4.2 In blue, distribution of p-values for the test that the spearman correlation between MI and relative evolutionary rate of each branch is significantly different from zero. In orange, the same test, but MI values are shuffled randomly across phylogenetic branches.

After correcting for multiple testing, we identified 78 genes showing a significantly higher or lower evolutionary rate correlated with MI (q-value < 0.1, supplementary table 4.2). 76 of these correlations were positive, indicating a higher evolutionary rate in placental species, and only two genes showed a negative correlation. The near-absence of slowly evolving genes in placental species may be due to the relatively short branches in the phylogeny: if the average number of mutations per branch is already low, the power to identify genes evolving at lower than average rates may be limited. Because of this, we focused on genes showing accelerated evolution in placental lineages for subsequent analyses.

GO enrichment analysis of accelerated genes revealed an enrichment of genes involved in metabolite transport. Overrepresented GO terms included “carboxylic acid transport” and “base amino acid transport” (supplementary table 4.3). For cellular components, the gene set was enriched in vesicle-located genes.

The *slc7a7* gene is an example of a transporter gene showing evidence of accelerated evolution in placental poeciliids. This gene codes for an amino acid transporter and is involved in nitric oxide synthesis in human umbilical vein (Arancibia-Garavilla et al. 2003). Almost all placental poeciliids show a faster evolutionary rate than expected for this gene, with *Heterandria formosa* and *Poeciliopsis prolifica* showing exceptionally high evolutionary rates (figure 4.3). These results suggest that the evolution of the placenta in the Poeciliidae is associated with consistent changes in nutrient transport systems. Four more examples of genes showing evidence of accelerated evolution in placental poeciliids are shown in supplementary figure 4.3.

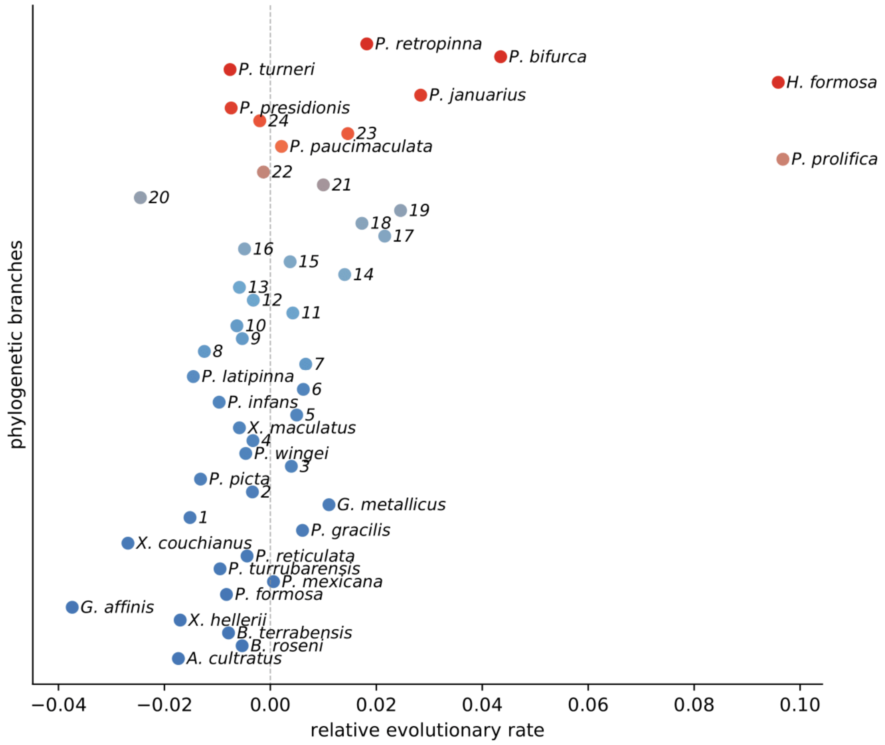

Figure 4.3 Relative evolutionary rate of the slc7a7 gene in all branches of the investigated phylogeny, sorted from high to low MI. The relative rate is defined as the deviation from the expected evolutionary rate, given branch- and gene-specific average rates. Points labelled 1 to 24 represent ancestral branches in the phylogeny, which have estimated MI values based

on ancestral trade reconstruction. See supplementary figure 4.2 for the poeciliid phylogeny with labelled ancestral branches corresponding to these points.

To gain more insight in the expression of our candidate genes in placental tissue, we used a previously published RNA-seq dataset of placental tissue of *Poeciliopsis retropinna* to see whether our candidate genes are expressed within placental tissue of this species (Guernsey et al. 2020). Using this dataset, we could confirm placental expression of 44 of 78 candidate genes (supplementary table 4.2). All amino acid transporters in our candidate gene list show expression in placental tissue of *P. retropinna*.

4.2.3 Convergent shifts in evolutionary rate are not due to positive selection

We tested the hypothesis that the accelerated evolutionary rate of our candidate genes was due to positive selection by quantifying the synonymous to non-synonymous mutation ratio (dN/dS, see methods). After correction for multiple testing, five out of 76 placenta-accelerated genes show evidence of positive selection in at least one branch leading to a placental species in the form of an elevated dN/dS mutation ratio (supplementary table 4.2). This is not a significantly higher proportion compared to the genome-wide proportion of genes evolving under positive selection in at least one of the placental branches (665 out of 14468 genes, p=0.41, Fisher's exact test). Additionally, evidence for positive selection was only apparent in one placental branch in the phylogeny for these five genes. A comparison of dN and dS between phylogenetic branches leading to placental species and other branches shows that for the majority of placenta-accelerated genes, both dN and dS are higher for placental branches than for other branches (supplementary table 4.2). This leads to an increased evolutionary rate while not necessarily increasing the dN/dS ratio.

Another hypothesis is that positive selection manifests on other genes than those for which we observe an accelerated evolutionary rate among placental species, but is still convergent among placental species. To test this hypothesis, we tested all the 14468 orthologous gene sets that were previously identified for positive selection using a branch-site model (see methods), with the phylogenetic branches leading to six highly placental poeciliids as foreground branches. As a control, we performed the same analysis using six non-placental species as foreground branches for which we do not expect any convergent molecular evolution. If placental species have more similar genes evolving under positive selection, the p-value distribution would be skewed towards lower p-values compared to the control group. However, the placental group did not show an excess of low p-values compared to the control group (supplementary figure 4.4), indicating that convergent placenta evolution does not manifest in a molecular signal that can be picked up by this test for positive selection. It seems that the genes evolving under positive selection show no similarity between different placental species, whereas changes that are captured by the evolutionary rate analysis do show a clear sign of convergence.

4.2.4 No excess of convergent amino acid substitutions in placental species compared to control groups

If certain amino acid substitutions are necessary for placenta evolution, these substitutions would have to be observed across all phylogenetic branches where a placenta has evolved. We tested whether we could observe an excess of these convergent substitutions among six placental branches in the phylogeny in the 14468 orthologous gene sets the we have identified across the Poeciliidae. For this, we compared the observed sequences of placental species with the inferred ancestral sequences. A convergent event was defined as an amino acid substitution at the same position in the protein for at least four placental species.

In total, we found 117 genes with one or more convergent amino acid substitutions in placental species. As a control, we performed the same analysis on 100 random combinations of six branches in the phylogeny, for which we do not expect convergent molecular evolution. We did not find an excess of convergent amino acid substitutions in placental species, compared to the control distribution (figure 4.4). This shows that although convergent amino acid substitutions happen somewhat frequently, convergent amino acid substitutions linked to convergent phenotypic evolution does not happen at a frequency that allows it to be detected in these species, if it occurs at all.

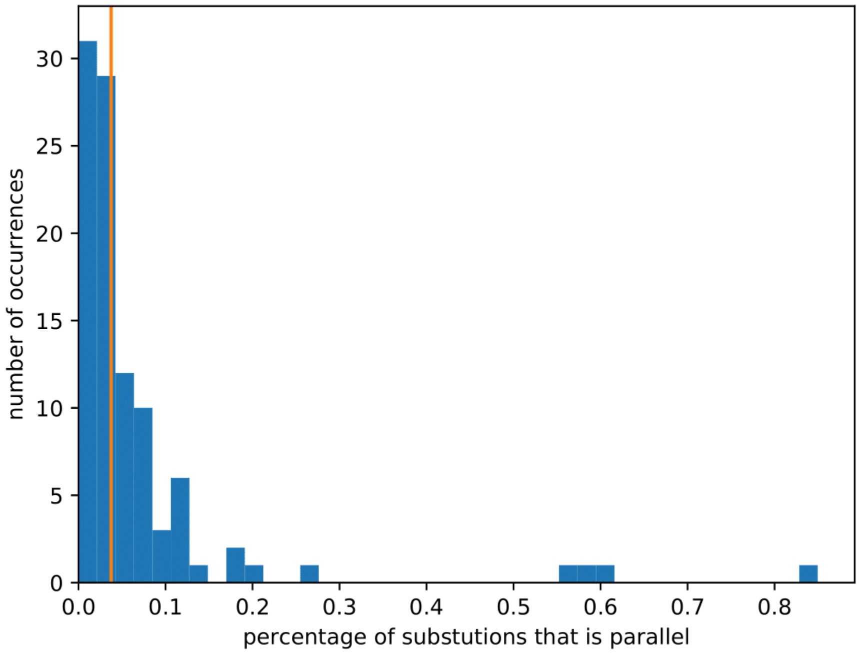

Figure 4.4 Percentage of convergent amino acid substitutions compared with the total number of substitutions in hundred random sets of six branches in the poeciliid phylogeny. Orange line: percentage of convergent amino acid substitutions compared with the total number of substitutions in the branches leading to six extensive matrotrophs in the Poeciliidae. For each test for convergent substitutions, we examined 14,468 genes.

4.2.5 Convergent differences in the conservation of non-coding elements between placental and non-placental poeciliid species

To also test for convergent changes outside protein-coding regions, we generated two multi-genome alignments: one with six placental poeciliids (MI > 10), and one with six non-placental poeciliids (MI < 1). To gain insight in more ancient conservation in our study system, we included a third alignment with five teleost fish species, four of which are outside of the family Poeciliidae (see methods for species). For all of these alignments, conserved elements were called independently using PhastCons (Hubisz et al. 2010). All of these conserved element predictions were aligned to the *P. retropinna* genome, allowing for a direct comparison of conserved elements in the different groups. Then, we identified conserved elements for placental poeciliids that showed no trace of conservation in the non-placental species, as well as conserved elements for non-placental poeciliids that showed no trace of conservation in placental species. This means that for these elements, genomic sequences are identical or almost identical across all species in one group (average nucleotide diversity of 0.11 substitutions per site), while for the other group there are a substantial number of mutations in these regions (average nucleotide diversity of 0.43 substitutions per site). This does not necessarily mean that these mutations have to occur in all species for the non-conserved group: a large number of mutations in the majority of the species will still lead to a low conservation score, even if the remaining species show high similarity to the conserved group. In total we found 76 of these Differentially Conserved Elements (DCEs). Fifty of these elements were more conserved in the placental species, and 26 were more conserved in the non-placental species. Our DCEs were highly enriched for simple sequence repeats (SSRs), especially GATA repeats: out of the 76 identified elements, 53 contained an SSR, and in 31 cases this was a GATA repeat. The proportion of DCEs containing a SSR is significantly higher than expected given the proportion of conserved elements containing a SSR across the whole genome (p=1.26e-85, Fisher's exact test).

To infer which genes these DCEs potentially regulate, we extracted each gene from the *P. retropinna* genome that lies directly downstream from a DCE (maximum distance 50kb). Additionally, we extracted genes for which a DCE lies in an intron or within the UTR of this gene. This resulted in a set of 85 genes that are potentially differently regulated in placental poeciliids compared to their non-placental counterparts (supplementary table 4.5). The Gene Ontology (GO) enrichment analysis revealed that this gene set was enriched for genes involved in several developmental processes, such as "anatomical structure development" or "system development" (supplementary table 4.6). By contrast, a control analysis consisting of the same workflow to test for differences between two sets of non-placental poeciliid species resulted in only 22 potentially differently regulated genes between these two groups of species, for which no significantly enriched GO terms were found (supplementary table 4.7). To test for a possible influence of different conserved element density across the genome, we also tested 25 sets of 76 random conserved elements for GO term enrichments, using the same strategy. This yielded a few enrichments for some of these sets, but none of them had such a clear overrepresentation of developmental genes as our candidate gene set (supplementary table 4.8).

A notable example of a difference in gene regulation between placental and non-placental species is found within a cluster of four ultra-conserved genes (*foxa2*, *pax1*, *nkx2-2a*, *nkx2-4*) that play a role in embryonic development. Synteny of this gene cluster is conserved across both teleost fish and mammals. A DCE was found between the *pax1* and *nkx2-2a* genes, consisting of a GATA simple repeat that was found in all placental species, but is not conserved in non-placental species (figure 4.5). The absence in conservation of this element in both the non-placental species, as well as the ancient conservation track suggests that this element has emerged in all placental species, although the partial presence of this element in some non-placental species indicates that this element did probably not emerge in the individual branches leading to placental species, but rather emerged deeper in the Poeciliid phylogeny and subsequently acquired mutations only in non-placental species. It has been shown that GATA repeat elements act as regulatory boundaries in both human and fruit fly, and a similar mechanism seems to be present in the Poeciliidae (Kumar et al. 2013).

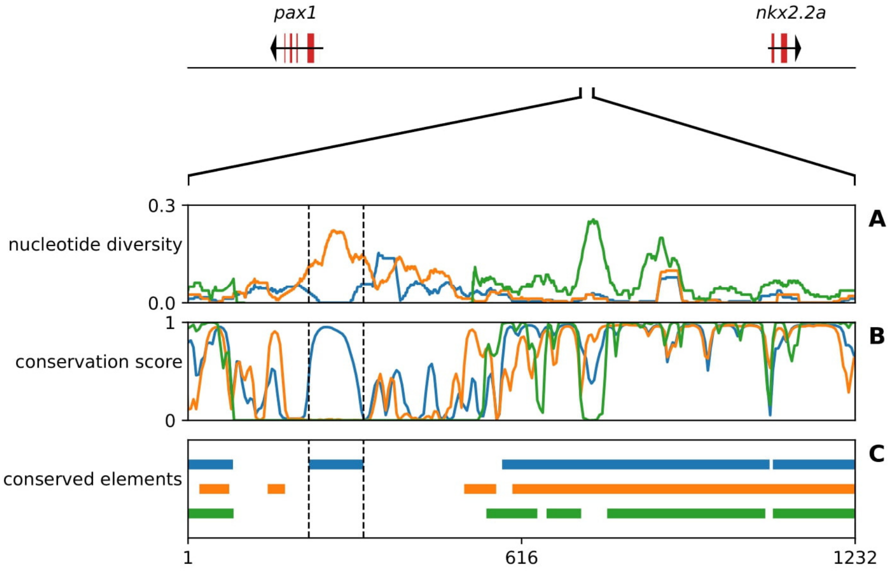

Figure 4.5 A placenta-specific boundary element near an ancient conserved element. Color codes: blue: placental poeciliids alignment; orange: non-placental poeciliids alignment; green: “ancient” teleost alignment. (A) Average nucleotide diversity for each alignment, 20-bp sliding window. (B) Per-base conservation scores, representing the estimated likelihood of the base belonging to a conserved element according to PhastCons. (C) Conserved elements predicted by PhastCons.

4.3 Discussion

Our results support the hypothesis that similar genomic changes underlie the repeated evolutionary origins of the placenta in the Poeciliidae. We find convergence in evolutionary rate shifts of protein-coding genes for placental species, as well as in differential conservation of non-coding elements for placental species as compared to their non-placental relatives. In both cases, we find a significantly higher incidence of these events compared to control analyses.

Our phylogenetic analysis based on a maximum likelihood strategy revealed a phylogeny that is identical to previous work (Pollux et al. 2014). However, some of the basal branches showed sub-optimal bootstrapping support. An alternative analysis based on a coalescent method resulted in a tree topology that is somewhat different at these basal nodes (supplementary figure 4.5). It seems that the earliest splits in the poeciliid phylogeny are hard to solve. However, running the downstream analyses on 100 random bootstrap trees showed that these slight differences in topology have very little effect on the results of the evolutionary rate analysis.

Many of the genes showing accelerated evolution in placental poeciliids are involved in vesicle functioning. This result is in agreement with a recent study showing that the placenta in the genus *Poeciliopsis* is based on a secretory system (Guernsey et al. 2020), a mechanism that likely applies to other genera in the Poeciliidae given the morphological similarities in placenta structure and development. The presence of several plasma membrane transporters in our accelerated gene set, such as the amino acid transporters *slc3a2*, *slc7a7* and *slc38a3*, suggests that additionally some nutrients are secreted through transporter proteins instead of vesicles. Investigation of a previously published RNA-seq dataset of placental tissue of *P. retropinna* (Guernsey et al. 2020) shows that these genes are indeed expressed in placental tissue of this species (supplementary table 4.2).

Despite the clear signature of convergence in the evolutionary rate of these genes, we did not find more genes evolving under positive selection in placental species than would be expected by chance. Prior studies employing an evolutionary rate approach to find convergent evolution also found only a small proportion of their candidate genes to evolve under positive selection (Chikina et al. 2016; Partha et al. 2017). An increase in evolutionary rate without evidence for positive selection is usually attributed to relaxation from evolutionary constraint. In our case this seems unlikely, as many of our candidate genes have vital functions in all fish species, and placenta evolution would presumably not affect this. A possible explanation for this discrepancy could be low power of tests for positive selection when selection is weak and phylogenetic branches are relatively short, as is the case for our phylogeny (Gharib and Robinson-Rechavi 2013). Alternatively, the observed changes in amino acid residues could be a consequence of positive selection acting outside of the coding region, with the increase in evolutionary rate inside the coding region being due to linkage disequilibrium rather than positive selection on certain amino acid residues. Further research, for instance on population genomic data, may be able to detect these selective forces.

The observed changes in the size and location of conserved elements between placental and non-placental poeciliids suggest convergent regulatory change in these species. However, the conservation differences between placental and non-placental species in our DCEs are not of the nature that the element is completely absent in the non-conserved group. Often, the element is still present or partially present in some of the species of the non-conserved group, but has acquired several mutations for others. This indicates that these elements likely did not emerge in the phylogenetic branch leading to highly placental species, but deeper in the phylogeny of the Poeciliidae. Then, the element was only conserved in either placental or non-placental species, while acquiring mutations in the other group. This suggests that some preconditions for placenta evolution may have already taken place before the evolution of a complex placenta in this family, which is not surprising given the continuous nature of the trait.

The overrepresentation of GATA repeats in our DCEs suggests a role of these elements in the gene regulation of these species. GATA repeats are known to function as enhancer blockers in both human and fruit fly (Kumar et al. 2013), can modulate promotor activity (Krishnan et al. 2017), and are associated with chromatin structure (Subramanian et al. 2003). GATA elements appear to have a specific length distribution, with GATA10-12 being the most abundant in humans (Kumar et al. 2010). In the Poeciliidae, we generally observe a prevalence of GATA20-40, longer than reported for other vertebrates (Kumar et al. 2010) (supplementary figure 4.6). The length of these repeats further suggests a functional significance, as without selection maintaining their length these would be highly unstable (Kim and Mirkin 2013). Indeed, other four-base SSRs do not show a preference for longer repeats (supplementary figure 4.7). Besides this apparent functional significance, SSRs are exceptionally variable, having mutation rates that are orders of magnitude higher than non-repetitive DNA (Gemayel et al. 2012). Because of this combination of having a functional role and being highly variable, GATA repeats could provide a potential mechanism for placenta evolution in the Poeciliidae by changing the regulation of developmental genes.

Finally, our results show that the evolution of the placenta in the Poeciliidae is accompanied by changes in both protein-coding and regulatory regions, suggesting that genomic changes in both categories are important for complex trait evolution. Notably, mutations in these two categories seem to be associated with different biological processes: protein-coding genes that show an evolutionary rate that correlates well with MI are mainly involved in metabolism and transport, while differentially conserved non-coding elements are mostly associated with development. This observed duality seems logical given that the amino acid sequences of developmental genes are usually highly conserved because they often act on many targets, making changes in the ensuing protein likely harmful due to pleiotropic effects. Although our study focuses on placenta evolution, the paradigm of protein-coding change in metabolic genes and regulatory change around developmental genes has been observed before and may be applicable to complex trait evolution in general (Partha et al. 2017; Sackton et al. 2019). As the genomes of more species become available this hypothesis can readily be tested in other evolutionary models.

4.4 Methods

4.4.1 Genome assemblies

15 genomes were assembled for this study, and 12 publicly available genome assemblies were included (supplementary table 4.9). Short-read assemblies were assembled from 30-50X coverage of 150bp paired-end Illumina reads using SPAdes 3.13.0 (Bankevich et al. 2012) with default settings, using the assembly generated with a k-mer size of 77. This is the default k-mer size for 150bp reads. Following assembly, contigs corresponding to heterozygous sections of the genome were removed using redundans v0.14a (Pryszcz and Gabaldón 2016), using settings --usebwa and --nogapclosing. For the *Phalloptychus januarius* assembly, the genome was assembled from 20X coverage of Oxford Nanopore long-reads (read N50 8.7kb) using Flye version 2.5 (Kolmogorov et al. 2019) with default settings and setting the estimated genome size to 600 Mbs. After assembly, the bases were polished by

mapping 30X coverage of 150bp paired-end Illumina reads to the assembly using BWA mem 0.7.17 (Li and Durbin 2009), before consensus calling with the wtdbg 2.5 consensus module (Ruan and Li 2019) with default settings.

4.4.2 Collecting orthologous genes

For assemblies of reference quality (see supplementary table 4.9), predicted protein sets were collected based on their annotations and analysed for orthology using ProteinOrtho v5.16b (Lechner et al. 2011) using default settings and blastp as the used program for alignment. Genes that displayed 1:1 orthology across all species were used for further analysis. Genes that displayed 1:1 orthology but were missing in one species were also used. This resulted in a set of 15,305 orthologous genes. Subsequently, the coding sequences of these genes were recovered from all short-read assemblies by aligning the coding sequence of the closest relative with a reference-quality genome to the assembly using exonerate version 2.2.0 (Slater and Birney 2005), using the cdna2genome model. Considering a 1:1 orthology in 12 reference genomes, a full-length match of the coding sequence in a single contig of the respective short-read assembly was assumed to be the 1:1 orthologous gene in this species too. Sequences with a premature stop codon were removed from the database. We continued analysis on genes of which we could recover the full coding sequence in at least three placental species, which was the case for 14,468 genes.

4.4.3 Construction of a molecular phylogeny

For the construction of a molecular phylogeny, we reconstructed the complete mitochondrial genome of all investigated species using MITObim 1.9.1 (Hahn et al. 2013) using settings --mismatch 1, -start 1 and -end 30, using the published mitochondrion of *P. reticulata* as reference. In addition, we made codon alignments of all orthologous genes recovered from all investigated species, as well as the non-poeciliid *Oryzias latipes* (1010 genes). Columns with gaps were removed from this alignment using trimAl v1.4 (Capella-Gutiérrez et al. 2009), using the -nogaps flag. The resulting alignments were concatenated into a ‘supermatrix’ alignment with a length of about 1.1 Mb. With this alignment, the phylogeny was reconstructed using RAxML 8.2.9 with the GTR+GAMMA model (Stamatakis 2014). *O. latipes* was used as an outgroup to root the tree. The RAxML analysis was done with eight partitions: three partitions for the three codon positions in both the mitochondrial coding sequence and the nuclear genes, one partition for mitochondrial non-coding RNA (tRNA and rRNA), and one partition for mitochondrial non-coding DNA (D-loop and some very small segments).

As an alternative analysis, we generated a phylogeny using a coalescent method. We made gene trees for each of the 1010 previously mentioned genes, as well as the mitochondrial genes using RAxML 8.2.9 with the GTR+GAMMA model (Stamatakis 2014). Then, we combined these gene trees into a species tree using ASTRAL v5.7.1, using default settings (Zhang et al. 2017). As ASTRAL does not estimate terminal branch lengths, branch lengths for the resulting phylogeny were estimated using RAxML 8.2.9 using option "-f e" for branch length estimation for a given topology.

4.4.4 Evolutionary rate analysis

Evolutionary rate analysis was performed as in Chikina et al (Chikina et al. 2016), slightly modified to test for a correlation with MI instead of a difference between two discrete classes. Amino acid alignments of previously identified orthologous genes were made using mafft v7.402 (Katoh et al. 2002). For each alignment, branch lengths were estimated for each branch across the reconstructed phylogeny using the AAML program of the PAML package (Yang 2007), using an empirical substitution model (Whelan and Goldman 2001). These raw branch lengths were converted into relative rates of evolution by normalizing for both the average rate of evolution of the investigated gene across all branches as well as the average rate of evolution of all genes within the investigated branch (as in Sato et al (Sato et al. 2005)). A resulting relative rate that is higher than zero corresponds to a gene that evolves faster than expected in the investigated branch, while a relative rate below zero corresponds to a gene that evolves slower than expected in the investigated branch. The relative rates were then used to test for the hypothesis that a gene evolves with a relative rate that correlates with the Matrotrophy Index (MI) - a proxy for placental complexity - using a Spearman ranked correlation test. For terminal notes in the phylogeny, this MI value was taken from Pollux et al., 2014 (Pollux et al. 2014). For ancestral nodes, the MI value was estimated using the phytools R package (Revell 2012). As a control, the same analysis was performed on the same dataset, but with MI labels for each node randomly shuffled across the phylogeny. On a genome-wide scale, the hypothesis was tested that more genes show an evolutionary rate that correlates well with MI than expected by comparing the case and control p-value distributions using the Kolmogorov-Smirnov test. To find candidate genes, correction for multiple testing was performed using the q-value method, with a threshold value of q=0.1 (Dabney et al. 2010).

Additionally, to assess robustness of results we repeated the ancestral trait reconstruction and evolutionary rate analysis on 100 random bootstrap trees (supplementary table 4.10). We confirmed our candidates for each of the 100 trees, and added the amount of trees that support our candidates in supplementary table 4.2.

4.4.5 Comparative non-coding elements analysis

To compare conserved non-coding elements between placental and non-placental poeciliids, two multi-genome alignments were made: one alignment consisting of the six placental species that have the highest MI values in the family (*Poeciliopsis retropinna, Poeciliopsis presidionis, Poeciliopsis turneri, Phalloptychus januarius, Heterandria formosa, Poecilia bifurca*), and one with six non-placental species that were chosen to follow a similar topology on the phylogeny as for the placental species (*Alfaro cultratus*, *Poeciliopsis turrubarensis, Poeciliopsis gracilis, Poeciliopsis infans, Poecilia picta, Xiphophorus hellerii*). A third multi-alignment consisting of five teleost fish (*Poeciliopsis retropinna*, *Oryzias latipes*, *Oreochromis niloticus*, *Gasterosteus aculeatus*, *Danio rerio*) was used to get insight on more ancient conservation, but it was not used for differential conservation analysis. The multi alignments were made by a custom pipeline (available on https:/ /git.wageningenur.nl/kruis015/whole_genome_alignment). All genomes were aligned to the chosen reference genome (*P. retropinna*) using the pairwise genome aligner MUMmer 4.0.0 (Marçais et al. 2018). After pairwise alignment, overlapping alignment blocks were merged based on reference coordinates and locally re-aligned using mafft v7.402(Katoh et al. 2002), yielding a reference-based multi-genome alignment. For both multi-genome alignments, conserved elements and base-specific conservation scores were called using PhastCons v1.5 (Hubisz et al. 2010). The non-conserved model was based on fourfold degenerate sites extracted from the respective alignment. For the non-placental alignment, the sequence of the placental reference genome (*P. retropinna*) was not used for predicting conservation by using the --not-informative option of the PhastCons program. In this way, base-wise conservation scores between the placental and the non-placental multi-alignment could be compared within the same genome, without using this genome for both predictions. After prediction of conserved elements, regions with a large difference in conservation between placental and non-placental species were extracted from the reference genome. The requirements for this were (1) the predicted element based on one multi alignment should not overlap with one based on the other multi alignment, (2) the mean difference in conservation scores across all bases of the predicted element should be at least 0.75, (3) there should be a significant difference in base-wise conservation scores between the two predictions based on a permutation test. As a control, the same analysis was performed, but instead of extracting differentially conserved elements between a placental and a non-placental set of species, differences between two non-placental sets of species were extracted. For this, the six non-placental poeciliid multi genome alignment as mentioned before was used and compared to a multi genome alignment of six other non-placental poeciliids (*Brachyrhaphis roseni*, *Gambusia affinis*, *Poecilia gilii*, *Poecilia mexicana*, *Poecilia reticulata*, *Xiphophorus maculatus*).

4.4.6 Simple Sequence Repeat analysis

To investigate the presence of Simple Sequence Repeats (SSRs) in our conserved elements, we identified SSRs across the genome of *P. retropinna* using MISA v1.0 (Thiel et al. 2003). The minimum number of repeated elements for identification of a SSR was given as ten repeats for an element size of one, six repeats for an element size of two, and five repeats for an element size of three or more.

4.4.7 Gene Ontology enrichment analysis

Gene Ontology (GO) enrichment tests and network analysis were performed using STRINGdb (database version 11.0)(Szklarczyk et al. 2016). For the evolutionary rate analysis, the predicted protein sequences of genes evolving at significantly different rates in placental species were extracted from the *P. retropinna* genome. The STRING database was searched for the human orthologs of these sequences, followed by manual curation when multiple candidates were presented. Subsequently, GO enrichment tests and network analysis were performed using the STRING web application (database version 11.0).

To find genes potentially regulated by our candidate DCEs, the first gene for which the element lies upstream was selected on both sides of the element as potentially regulated gene. If the first gene next to the element was not in the orientation so that the element lies upstream of the gene, the gene was not selected. Also, the gene was not selected if it was further than 50kb away from the element. The identified genes were subjected to the same analysis in STRINGdb as done for the genes identified in the evolutionary rate analysis.

4.4.8 Detecting positive selection

We performed tests for positive selection on the 14,468 orthologous gene sets that were previously identified. For each gene, a codon alignment was made using PRANK v.170427 (Löytynoja 2014), using options -codon and -F. These alignments were used to test for positive selection with the codeml program that is part of PAML 4.9 (Yang 2007).

To test for positive selection that may arise during the evolution of the placenta, we used the so-called branch-site model to test for positive selection for every branch leading to all placental species (*P. retropinna, P. paucimaculata, P. presidionis, P. turneri, P. prolifica, P. januarius, H. formosa, P. bifurca*). Each gene was tested for every placental branch separately. The hypothesis that genes evolving under positive selection in placental species are overrepresented in the set of genes that show accelerated evolution in placental species was tested using Fisher's exact test.

Additionally, convergence among positive selection when evolving a placenta was tested by using the same branch-site model, but now using the phylogenetic branches leading to six highly placental poeciliids (*P. retropinna, P. presidionis, P. turneri, P. januarius, H. formosa, P. bifurca*) as foreground branches. As a control, the same analysis was performed to a group of six non-placental species (*A. cultratus, P. turrubarensis, P. gracilis, P. infans, P. picta, X. hellerii*) for which we do not expect genomic convergence. We then compared the distribution of p-values between the placental group and the control group to see if there is a consistent enrichment of low p-values when using the placental species as foreground branches using the Kolmogorov–Smirnov test.

4.4.9 Detecting convergent amino acid substitutions

To test whether placental species disproportionally show convergent amino acid substitutions, we reconstructed ancestral sequences of each of the 14,468 orthologous gene sets that were previously identified with the AAML program that is part of PAML 4.9 (Yang 2007) , using an empirical amino acid substitution matrix (Whelan and Goldman 2001). For six highly placental species (*P. retropinna, P. presidionis, P. turneri, P. januarius, H. formosa, P. bifurca*) we compared the observed amino acid sequence with that of its closest ancestor. However, we took an exception for the ancestor of *P. presidionis* and *P. turneri*, as these species represent a single origin of placentation in the Poeciliidae, and their common ancestor is hypothesized to have a placenta a well. Therefore, we compared both the sequences of *P. presidionis* and *P. turneri* with the common ancestor of these two species and the non-placental *P. gracilis*. In these comparisons, amino acid substitutions that occur on the same position in at least four out of six comparisons were identified. These amino acid substitutions were noted as potential convergent events. Then, the same analysis was performed for 100 random combinations of six species, to get a background distribution of convergent amino acid substitutions when no morphological convergence is apparent.

4.4.10 Placental expression analysis

To confirm expression of candidate genes in placental tissue, we downloaded RNA-seq data gathered from follicular epithelium of the placental *P. retropinna* (Guernsey et al. 2020). This data was mapped to the *P. retropinna* genome using HISAT2 v2.1.0 (Kim et al. 2019), using default settings. To detect expression of candidate genes, the coverage of mapped reads was determined for all exonic positions of candidate genes using samtools depth (Li et al. 2009). An average read coverage of 2x across all exons was used as the cutoff for gene expression.

4.5 Additional files

The online version of this article (10.1093/molbev/msab057) contains supplementary material, which is available to authorized users. Additionally, all supplementary material referenced to in this thesis is available on the Zenodo database (10.5281/zenodo.5647272).

4.6 Acknowledgements

This work was supported by a WIAS research grant from Wageningen Institute of Animal Sciences, awarded to HJM & BJAP, and a VIDI grant from the Netherlands Organisation for Scientific Research (864.14.008) awarded to BJAP.
All animal-derived biological materials used this study were obtained according to local ethical regulations. Samples were collected and exported by B.J.A.P. following the regulations of the Costa Rican government under permit number SINAC-CUS-PI-R-005-2017 or obtained from D.N.R. (University of California Riverside, USA).

4.7 Authors’ contributions

Biological material for sequencing was obtained from DNR and BJAP. HVK and RN performed the sequencing. HVK performed analyses and wrote the draft manuscript, supervised by HJM and BJAP. RN, DNR, MAMG, HJM and BJAP provided recommendations to improve the manuscript. All authors approved the final manuscript.

4.8 Competing interests

The authors declare that they have no competing interests.

5

Genetic losses rather than duplications are associated with placenta evolution in livebearing fish

Henri van Kruistum[1,2], Martien A. M. Groenen[1], Hendrik-Jan Megens[1,3] and Bart J.A. Pollux[2]

[1] Animal Breeding and Genomics group, Wageningen University, The Netherlands.
[2]Experimental Zoology group, Wageningen University, The Netherlands.
[3]Aquaculture and Fisheries group, Wageningen University, The Netherlands

Abstract

During vertebrate evolution, complex organs have evolved several times. It has been hypothesized that gene duplications can drive the evolution of new complex traits. However, testing this prediction consistently poses a challenge because the emergence of a new complex trait is relatively rare. Here, we test the hypothesis that gene duplications are associated with complex trait evolution in the livebearing fish family Poeciliidae. In this family, a placenta has evolved nine times independently, allowing for the study of genomic convergence in association with placenta evolution. We show that different instances of placenta evolution are not associated with gene duplications in the same genes, nor in genes with a similar biological function. However, we do find an association with placenta evolution and loss-of-function mutations within genes involved in certain developmental pathways, specifically the calcineurin-NFAT signaling cascade and the BMP signaling pathway. We hypothesize that removing a copy of some previously duplicated genes that are a result of the teleost-specific genome duplication can influence these developmental pathways in some tissues, while other tissues are not affected by the deletion, thereby allowing to circumvent deleterious pleiotropic effects that would normally accompany such a deletion.

5.1 Introduction

Studying the evolution of vertebrate morphological diversity is a major goal in evolutionary biology. As all species develop based on their genomic information, morphological variation should be encoded in the genome. This allows for studying the evolution of morphological diversity by studying genomic diversity. As the availability of genomics data increases, it becomes increasingly straightforward to associate simple phenotypic variation with genomic mutations. However, the association of genomic data with more complex evolutionary novelties is less clear-cut, as generally more than a few simple polymorphisms are necessary for the generation of complex traits, such as the evolution of an organ. Rather, complex trait evolution is thought to be associated with larger-scale variation, such as duplications or translocations of genes or regulatory elements (Prud'homme et al. 2007; Wagner 2008).

Even before the genomics era, gene duplications have been predicted to play a major role in the evolution of complex traits (Ohno 1970; Holland et al. 1994; Wagner 1994). A duplication of an essential gene would allow for more flexibility in the evolution of both coding and regulatory sequences, as deleterious pleiotropic effects would be circumvented when conserving one of the two copies. The process of gene duplication to relieve regulatory constraints was later observed in yeast (Hittinger and Carroll 2007). Since then, gene duplications have been proposed to play a role in the evolution of a variety of complex traits. A few examples of this are the evolution of C4 photosynthesis in plants (Bianconi et al. 2018), parasite-to-host adaptation in a fish parasite (Konczal et al. 2020), or centromere function in *Drosophila* (Ross et al. 2013). Although these reports show that gene duplications in certain situations can be associated with adaptive evolution, a causal link between gene duplications and complex trait evolution in general is hard to establish, as most of these traits evolve only once.

To reliably correlate gene duplications with evolutionary novelties, one would need a model in which a trait evolves multiple times independently in closely related species. Such a model can be found in the fish family Poeciliidae. The Poeciliidae are a family of livebearing fish that consist of about 275 species (Parenti 1981; Van Der Laan et al. 2014). Most species in this family are lecithotrophic, meaning that nutrients are supplied to the offspring before fertilization, in the form of egg yolk. However, some species have evolved a placenta and supply nutrients through this placenta after fertilization, a phenomenon known as matrotrophy. This placenta evolution has occurred nine times independently in this family (Pollux et al. 2009; Furness et al. 2019), which allows for the study of multiple occurrences of complex trait evolution in closely related species.

Although the genomic basis of this placenta evolution has been studied before (Van Kruistum et al. 2019; van Kruistum et al. 2020; van Kruistum et al. 2021), a genome-wide assessment of gene duplications and gene losses in the phylogenetic branches leading to placental species has, until now, been performed on one placental species only (van Kruistum et al. 2020). Therefore, it is currently unknown whether the structural variants found in this past study apply to other instances of placenta evolution as well. In this study we perform a genome-wide assessment of gene duplications in the genomes of ten Poeciliid species, of which three have a placenta. As a result of this repeated evolution, the importance of gene duplications can be assessed in multiple independent origins simultaneously. Using this dataset we test the prediction that gene duplications are playing a role in the evolution of complex traits: if gene duplications play a major role in placenta evolution, gene duplications in the same genes or genes involved in a similar function would be observed for each placental species, while being absent in closely related non-placental species.

Besides gene duplications, we assess the occurrence of gene loss in placental poeciliid species. This is interesting in the context of immune system evolution, as the evolution of the placenta is hypothesized to be associated with some form of immune suppression (Gobert and Lafaille 2012). In seahorse and pipefish, species that have evolved a brood pouch simultaneously lost their MHC II system (Roth et al. 2020). If the association between placenta evolution and immune gene loss is consistent, we should be able to detect immune gene loss in placental poeciliid species as well.

5.2 Results

To assess the occurrence of structural variation in the Poeciliidae, we used genome assemblies of ten Poeciliid species, three of which have a placenta (green branches in figure 5.1). Additionally, we used five outgroup species to compare structural variation also outside of the Poeciliidae (grey branches in figure 5.1).

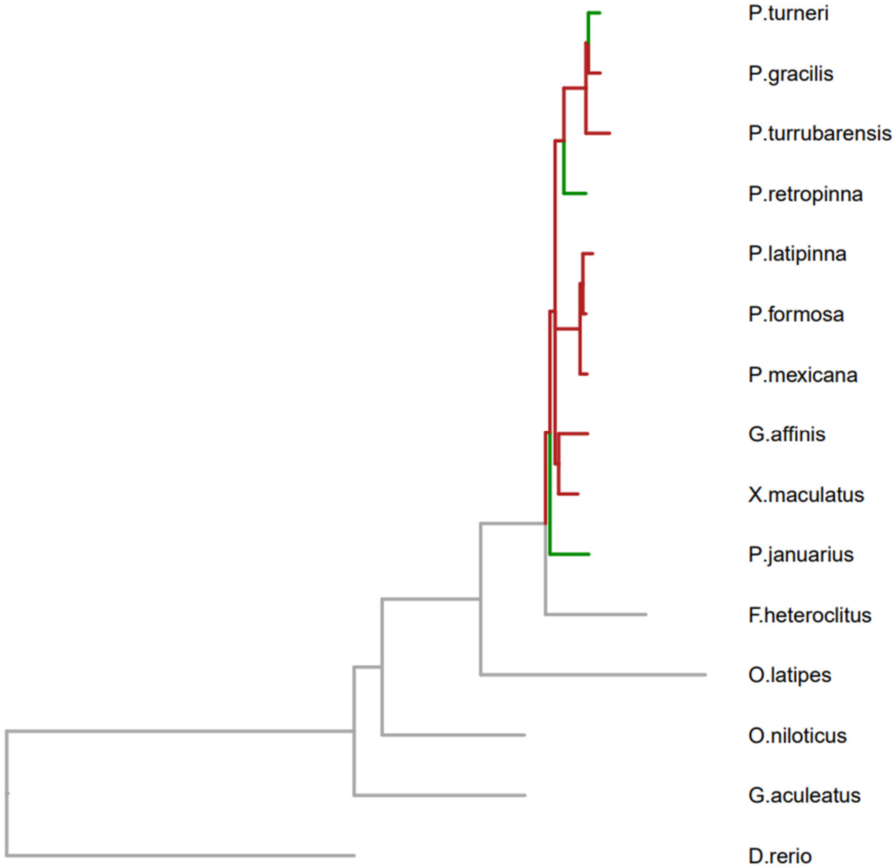

Figure 5.1 Molecular phylogeny of the investigated species. Red branches: non-placental livebearing species. Green branches: placental livebearing species. Grey branches: outgroup (egg-laying) species.

We applied a whole genome alignment based method to find gene duplications in several clades of the phylogeny. For the family Poeciliidae as a whole, we found a small number of duplicated segments that did not contain any genes. A much larger part of the outgroup genomes were deleted in the Poeciliidae, which can be seen in the large amount of deleted segments and genes (table 5.1). For each of the placental species, we found a small number of duplicated genomic segments that are not present in any of the non-placental species, and a somewhat larger number of deleted genomic segments that are still present in all non-placental species. However, almost all of these structural variants are only found in one placental species. This shows that there is no convergent evolution on the level of segmental duplications or deletions in the three placental species that we investigate.

Table 5.1 Number of duplicated and deleted genomic segments and the genes within these segments in all species of the Poeciliidae, the number of duplicated and deleted genomic segments observed in the three placental species, while being absent in any non-placental species, as well as the overlap between then duplications and deletions in the three placental species

Species group	Duplicated segments	Duplicated genes	Deleted segments	Deleted genes
All Poeciliidae	16	0	1248	28
P. januarius	72	5	265	13
P. retropinna	49	3	176	33
P. turneri	83	13	249	36
All three placental	1	0	1	0

Although this method exhaustively detects deleted and duplicated genomic segments, this does not capture all forms of gene loss, as in theory a mutation of a single nucleotide could render a gene inactive, while the genomic segment itself would not be lost. To detect these deleterious variants, we mapped all predicted proteins of the *Oryzias latipes* genome annotation to their approximate corresponding position in all poeciliid genomes, and checked for indications of deleterious mutations in the resulting alignment (stop-gain or frameshift mutations). A substantial number of *O. latipes* transcripts showed indication of loss-of-function in their poeciliid counterparts: we found 177 genes for which loss-of-function mutations were found in all poeciliid species (table 5.2). Even higher numbers were found for species-specific loss-of-function mutations in placental species. However, the number of genes that showed loss-of-function mutations in all placental species and were still functional in any non-placental species was not significantly higher than expected by chance (p = 0.091, permutation test, figure 5.2). This shows that placental species do not show genomic convergence in regard to gene loss.

Table 5.2 Number of genes for which a loss-of-function mutation was observed, relative to the outgroup species, for all species of the Poeciliidae, and the number of genes for which a loss-of-function mutation was observed in the three placental species, while a fully intact coding sequence was still found in all non-placental species.

Species group	Genes with loss-of-function mutation
All Poeciliidae	177
P. januarius	585
P. retropinna	533
P. turneri	501
All three placental	8

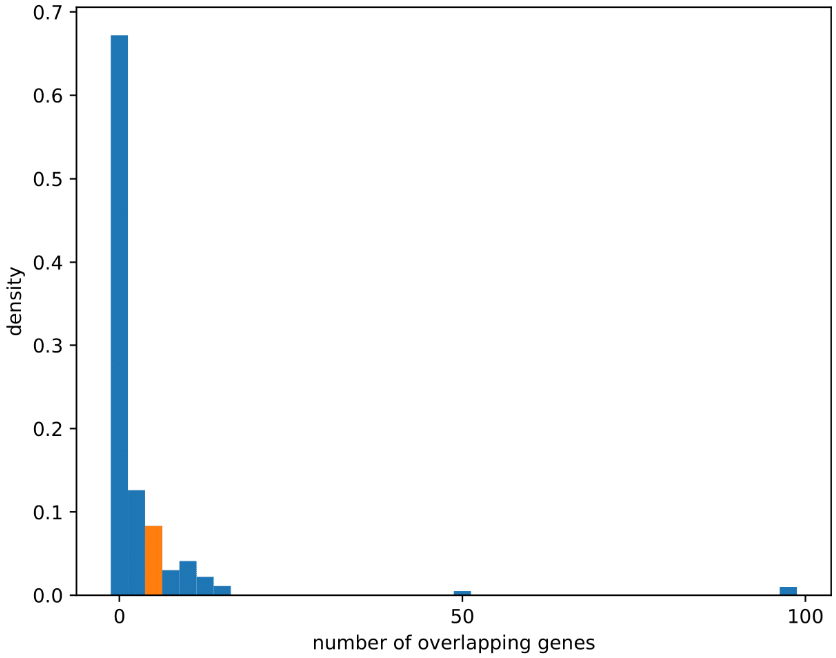

Figure 5.2 in blue: histogram of the distribution of the number of overlapping genes between three sets of randomly selected genes. In orange: bin of the histogram containing the overlap of three sets of genes with a loss-of-function mutation in placental species.

An alternative hypothesis is that although the placental species show no convergence in structural variants on the gene level, the duplicated and deleted genes are involved in a similar function (functional convergence). To test this hypothesis, we downloaded Gene Ontology (GO) annotations for genes inside structural variants or genes having loss-of-function mutations and scanned the candidate gene sets for overlapping ontology terms. GO terms that were present at least two times in each candidate gene set were tested for a higher overlap than expected by chance between the candidate gene sets of different placental species by sampling random gene sets of the same size as our candidate gene sets, and making a distribution for the overlap occurrence. For genes inside structural variants, we see for some GO terms a slightly greater than expected degree of overlap between the candidate gene sets, but these results were not statistically significant after correction for multiple testing (supplementary table 5.1). From this we can conclude that we have no evidence that the observed structural variants in different placental species show convergence in molecular function. However, for genes showing loss-of-function mutations we see a significantly higher co-occurrence of several GO terms than expected by chance (figure 5.3, supplementary table 5.2). This shows that although loss-of-function mutations in different placental species do not coincide on the same gene more than expected by chance, they do coincide in genes having a similar function.

Of the 73 GO terms that co-occur more than expected by chance in gene losses in placental species, many are related to neuron development (supplementary table 2). Additionally, we found evidence for co-occurrence of gene losses involved in signaling pathways, specifically the calcineurin-NFAT signaling cascade and the BMP signaling pathway (supplementary table 2).

Although this analysis shows that gene losses in placental species occur in genes with related functions, this does not distinguish whether this gene loss is associated with placenta evolution or with poeciliid evolution in general. To test this, we performed the same analysis on three non-placental species with similar phylogenetic topology (*Poeciliopsis gracilis*, *Poeciliopsis turrubarensis* and *Xiphophorus maculatus*). For this species trio, we found that 11 GO terms co-occurred more than expected by chance (supplementary table 5.3). Again, we found several neuron related GO terms in this set, suggesting that neuron related gene loss is prevalent across the poeciliid family. However, we did not find any loss of genes involved in the calcineurin-NFAT or BMP signaling cascades in these species, suggesting that gene loss in this pathway is restricted to placental species.

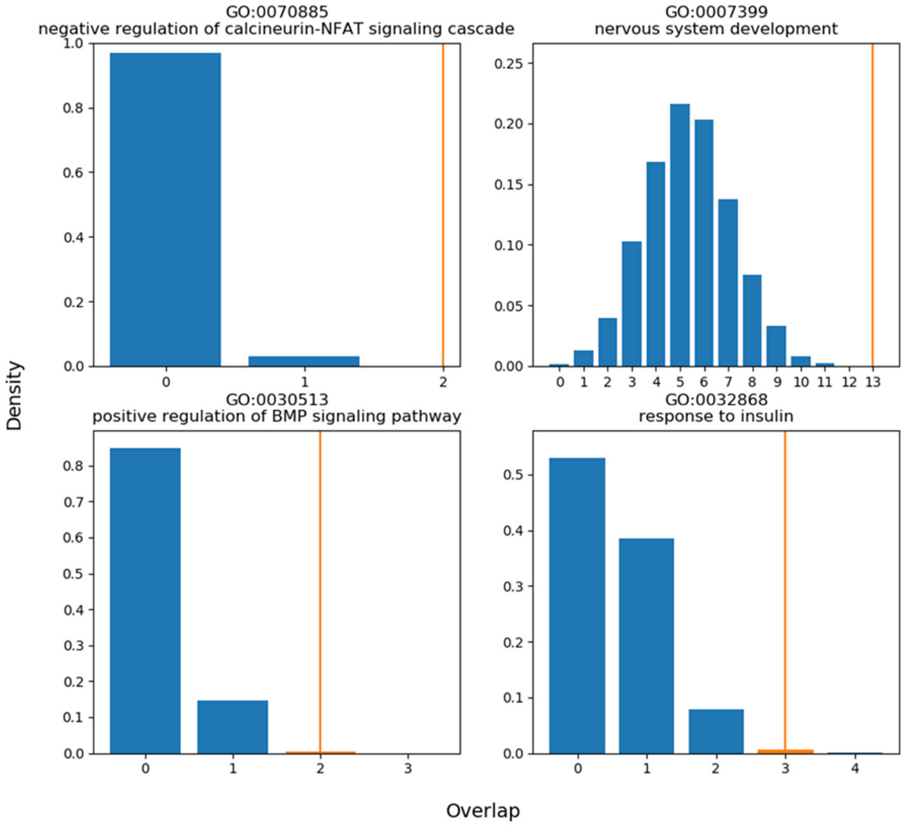

Figure 5.3 In blue: for four GO terms, probability distributions for the minimum number of genes of this GO term showing a loss-of-function mutation in all placental species, given the observed amount of genes showing a loss-of-function mutation randomly drawn from the complete pool of genes. In orange: for four GO terms, the observed minimum amount of genes of this GO term showing a loss-of-function mutation in all placental species.

5.3 Discussion

In this study, we tested the occurrence of gene duplications and deletion in relation to placenta evolution in poeciliid fish. Specifically, we aimed to test two hypotheses: (1) whether the evolution of the poeciliid placenta coincides with specific gene duplications that would explain the emergence of this new organ, and (2) whether the evolution of the poeciliid placenta coincides with gene loss of immune genes that would otherwise promote rejection of the embryo.

We did not find any evidence that confirms the first hypothesis. Although all three placental species showed some species-specific gene duplications when compared with the outgroups, none of these duplications were shared between all three placental species. Additionally, the duplicated genes were not involved in similar function more than would be expected by chance. The same results were found for duplications in non-coding regions, indicating that neither coding nor non-coding segmental duplications are involved in the evolution of the placenta. This suggests that no new genes are necessary for the poeciliid placenta to evolve, but instead repurposing of existing genes is sufficient.

We do find some evidence suggesting that placenta evolution in livebearing fish may be accompanied by gene loss of genes involved in similar functions. For instance, all three placental species showed loss of genes involved in the calcineurin-NFAT signaling cascade. Normally, this signaling cascade activates when Ca2+ enters the cell, activating calcineurin. Calcineurin dephosphorylates members of the NFAT protein family, which then move into the nucleus and activate transcription of downstream targets (Crabtree and Schreiber 2009). Interestingly, all genes that were found to contain loss-of-function mutations in placental species have been identified as an inhibitor of NFAT signaling in other species: both *mtor* and *dyrk2* phosphorylate NFAT proteins, preventing downstream activity (Gwack et al. 2006; Yang et al. 2008). The *myoz1* and *fhl2* genes inhibit NFAT signaling by inhibiting calcineurin activity, therefore inhibiting dephosphorylation of the NFAT proteins (Frey et al. 2008; Hojayev et al. 2012). This seems counterintuitive, as overexpression of this pathway in immune cells is associated with a more active immune system and higher inflammation levels in humans (Park et al. 2020), while the evolution of the placenta is expected to be associated with some form of immune suppression (Gobert and Lafaille 2012). However, based on this data we cannot gather in which cell types the inhibition of NFAT signaling is diminished, and the genomic targets of this signaling cascade are highly varied, depending on cell type (Crabtree and Schreiber 2009). For instance, NFAT activation in endothelial cells is essential for the development of new blood vessels, a phenotype that would be consistent with placenta evolution (Graef et al. 2001; Moccia et al. 2019). Further research is needed to identify the cell types in which NFAT signaling is affected to gain more insight in the phenotypic result of these deletions.

Another interesting developmental pathway in which placental species showed and excess of genes with a loss-of-function mutation is the BMP signaling pathway (GO:0030513). This signaling pathway is involved in the development of a wide range of tissues, such as bone, vascular and ovarian tissue (Wang et al. 2014). Unlike the deleted genes involved in the NFAT signaling pathway, the genes deleted in placental species that are involved in BMP signaling are not involved in one specific molecular function, and seem to be influencing BMP signaling in a wide variety of cell types (supplementary table 5.2). Deletions of genes within the BMP signaling pathways could have numerous consequences, as these pathways have broad functions in energy metabolism and development.

Concluding, we do not find any evidence for involvement of structural variants in placenta evolution in the "classical" sense, that is, gene duplications leading to neo- or sub-functionalization. We do find an association of placenta evolution with loss-of-function mutations in genes involved in several developmental pathways. Although some of these genes seem vital for proper functioning of these pathways, it has to be noted that because of the teleost-specific genome duplication (Glasauer and Neuhauss 2014), in most of these cases a second copy of the gene corresponding to the same human ortholog is still present, without evidence of functional loss. This suggests that the corresponding pathways will only be influenced in cell types that express the copy showing a loss-of-function mutation. Pruning the expression of previously duplicated genes in this way may allow for these developmental pathways to evolve in certain tissues without resulting in deleterious pleiotropic effects in other tissues.

5.4 Methods

5.4.1 Genome assembly of *Poeciliopsis gracilis* and *Poeciliopsis turneri*

To assemble the genomes of *P. gracilis* and *P. turneri*, a male specimen of both species was sacrificed using a lethal dose of MS-222. DNA was then isolated from fin tissue using the Qiagen Genomic-Tip 100/G DNA extraction kit. This DNA was then sequenced on an Oxford Nanopore MinION sequencer, yielding around 30X coverage for the *P. gracilis* individual, and 22X for the *P. turneri* individual. Subsequently, the sequencing data was assembled using Flye version 2.5 (Kolmogorov et al. 2019), using default settings and setting the estimated genome size to 600 Mb. The resulting assembly was polished using ~40X coverage of Illumina short reads from the same individual, by mapping the short reads to the assembly using BWA mem version 0.7.17 (Li and Durbin 2009), followed by consensus calling using the redbean version 2.5 consensus module (Ruan and Li 2020).

5.4.2 Whole genome alignment

A multi genome alignment of all 17 fish species investigated in this study was constructed with Cactus version 1.3.0 (Armstrong et al. 2020), using default settings. To construct the phylogenetic tree required for the input, we used the tree topology as determined in an earlier study (van Kruistum et al. 2021) for poeciliid species, and then adding the outgroups to this tree as determined by their taxonomy (from ncbi taxonomy). Then, branch lengths were estimated by aligning full mitochondrial genomes of all investigated species using mafft v7.402 (Katoh et al. 2002), and then estimating branch lengths using RAxML version 8.2.9 (Stamatakis 2014) using option '-f e' to estimate branch lengths given a tree topology.

5.4.3 Identifying deleted and duplicated genomic segments

Using the previously generated whole genome alignment, we identified duplicated and deleted genomic segments in the genomes of poeciliid species relative to the outgroup species. For this analysis, we used the genome of *Fundulus heteroclitus* as a reference, as this is the outgroup most closely related to the Poeciliidae. For each species, we extracted pairwise synteny blocks of this species to the *F. heteroclitus* genome using the halLiftover utility of the cactus tool, with option --outPSL to output synteny blocks. Then, we created a database where for each base of the *F. heteroclitus* genome, we counted the number of synteny blocks of each species overlapping this base. To then identify deleted or duplicated genomic segments, we searched for genomic segments with zero (for deletions) or more than one (for duplications) synteny blocks overlapping its bases for our species or species group of interest, while checking for one-on-one synteny with the other outgroup species.

5.4.4 Identifying loss-of-function mutations

To identify genes showing a loss-of-function mutation in poeciliid species, we used the *Oryzias latipes* genome as a reference, as this is the outgroup most closely related to the Poeciliidae that has a high-quality annotation. For each transcript in the *O. latipes* genome annotation, the orthologous genomic region for all poeciliid species was extracted using the halLiftover utility of the cactus tool. Then, the predicted protein sequence was aligned to this region using exonerate version 2.2.0 (Slater and Birney 2005), using the protein2genome model. The resulting alignment was then scanned for stop-gain or frameshift mutations for each species.

5.4.5 Gene Ontology enrichment tests

To test the hypothesis that duplicated or deleted genes in different placental species show a higher degree of congruence in biological function than expected by chance, we performed a permutation test. For each candidate gene, Gene Ontology (GO) annotations were downloaded from the gene ontology website. Then, for each GO term, the amount of genes being annotated with this go term was counted for each of the three placental species' candidate gene sets. We then defined the lowest observed count as the overlap value, e.g. the minimum amount of genes being annotated with this particular GO term being present in each candidate gene set. Then, a background distribution of overlap values for each GO term was constructed by taking 10000 random gene sets of identical sizes as our candidate gene sets, and calculating overlap values for each of these sets. The raw p-value is then defined as one minus the fraction of values in the overlap distribution lower than the observed value. Raw p-values were then corrected for multiple testing using the Benjamini-Hochberg procedure (Benjamini and Hochberg 1995).

5.5 Additional files

All supplementary material referenced to in this thesis are available on the Zenodo database (10.5281/zenodo.5647272).

5.6 Acknowledgements

Not Applicable

5.7 Author contributions

HvK performed the analyses and wrote the first version of the manuscript. MAMG, BJAP and HJM critically reviewed the manuscript for further improvement.

6

Allele-specific DNA methylation in the livebearing fish *Poeciliopsis gracilis*

Henri van Kruistum[1,2], Reindert Nijland[3], Tiffany R. Ernst[2], Ole Madsen[1], Martien A.M. Groenen[1], Bart J.A. Pollux[2] and Hendrik-Jan Megens[1,4]

[1]Animal Breeding and Genomics group, Wageningen University, The Netherlands. [2]Experimental Zoology group, Wageningen University, The Netherlands. [3]Marine Animal Ecology Group, Wageningen University, The Netherlands. [4]Aquaculture and Fisheries Group, Wageningen University, The Netherlands

Under review

Abstract

DNA methylation plays an important role in the regulation of gene expression. In a diploid organism, the two haplotypes are expected to have the same methylation state for most of the genome. However, differences in methylation between the two haplotypes can occur at certain genomic regions, which can result in allele-specific expression of nearby genes. Allele-specific methylation has been studied in mammals and flowering plants, but studies in other organisms are scarce. Here we look for allele-specific methylation in the genome of the livebearing fish *Poeciliopsis gracilis* by sequencing four individuals: two parents and two of their offspring. We developed a new approach to detect allele-specific methylation based on the Oxford Nanopore long read sequencing technology, and use a trio binning approach to link the alleles of sites with evidence for allele-specific methylation to their parent-of-origin. We found allele-specific methylation to be widespread across the genome of *Poeciliopsis gracilis*, depending on the individual affecting CpG sites around 824 to 3442 protein-coding genes. We show that although heterozygous positions at CpG sites can explain some of these observations, most CpG sites showing allele-specific methylation do not contain heterozygosity. We also show that allele-specific methylation in the offspring is not random in terms of parent-of-origin, with the methylated allele more often than expected originating from the parent of the opposite sex. Genes that are nearby or overlapping regions showing allele-specific methylation are significantly enriched for neurological function. Our results show that allele-specific methylation is widespread in the genome of *Poeciliopsis gracilis*, and that its pattern of inheritance is non-random. We hypothesize that these patterns of allele-specific methylation can evolve if different selective pressures exist for the two sexes on the same locus, and that sexual selection as observed throughout the genus *Poeciliopsis* can explain these results.

6.1 Introduction

DNA methylation is a common epigenetic mechanism in eukaryotes to control gene expression. DNA methylation refers to the addition of a methyl (CH_3) group to a nucleotide residue, often cytosine. These cytosine residues are methylated to form 5-methylcytosines, in vertebrates usually when the cytosine is followed by a guanine residue (i.e. a CpG site). Methylation of CpG sites within a gene can change its expression. When first discovered, methylation of CpG sites was predominantly associated with gene silencing (Vardimon et al. 1982). However, later it was discovered that DNA methylation inside the gene body could also enhance gene expression when combined with an unmethylated promoter region (Ball et al. 2009; Yang et al. 2014). Generally, methylation inside the promoter region is thought to be associated with gene silencing, while unmethylated CpG sites in the promoter correspond with actively transcribed genes (Razin and Cedar 1991; Illingworth and Bird 2009).

Although in a diploid organism most CpG sites are either symmetrically methylated or non-methylated at both chromosomal copies, some CpG sites are methylated on only one of the two alleles, a phenomenon known as allele-specific methylation (ASM). ASM can lead to skewed or even mono-allelic expression of nearby genes. The emergence of ASM in evolution can be easily explained in some contexts but has yet to be completely explained in others. There are several recognized mechanisms that may lead to ASM.

First, random epigenetic inactivation of one of the X chromosomes in diploid organisms with an XY sex-determination system can lead to ASM. X-chromosome inactivation refers to the transcriptional silencing of one X chromosome in XX females, preventing them from making twice as many gene products as XY males [6]. In eutherian mammals X-chromosome inactivation occurs early during embryonic development with both female X chromosomes having an equal probability of being silenced (Gartler and Riggs 1983). In this case, ASM is thought to act as a mechanism to avoid that the different number of X chromosomes between mammalian males and females affects viability (Heard et al. 1997).

Second, ASM may be mediated by SNPs inside CpG sites. When a heterozygous SNP occurs inside a CpG site, this site may be disrupted for one of two alleles, leading to ASM. A study by Shoemaker et al (2010) revealed SNP-mediated ASM in human cell lines of multiple tissues (Shoemaker et al. 2010). They showed that, depending on the cell line, 38-88% of ASM in humans could be explained by heterozygous SNPs inside CpG sites. When this happens inside promotor or enhancer regions, this can lead to allele-specific gene expression.

Third, genomic imprinting refers to the asymmetric silencing of one of the parental alleles through epigenetic mechanisms, such as ASM (Li et al. 1993b). The key feature of genomic imprinting is that ASM is not dependent on genomic variation as is SNP-mediated ASM but is instead dependent on the parent of origin of the allele. Several theories attempt to explain the emergence of genomic imprinting in evolution.

The genetic conflict theory (Haig 2000) proposes that a dispute over the allocation of nutrients between mother and her developing offspring is driving the asymmetric silencing of one of the parental alleles. Parent-offspring conflicts are shaped by opposing interests of the mother versus the father on how much resources are spent by each to benefit the offspring. Such conflicts are especially notable in placental mammals, where the mother provides all nutrients for early growth, which, if demand from the offspring is too high, may go at the expense of the female’s survival and future reproductive success. Therefore, genes that drive demand for nutrients during development need to be balanced in their expression, specifically paternally-derived fetal genes. For the male, it may be beneficial to increase the transfer of nutrients to the fetus, while for the mother it is important to balance investment with survival and future reproductive success. This conflict between maternal and paternal genes is thought to drive the silencing of one of the parental alleles of genes involved in nutrient transfer and embryonic growth (Haig 2000), with genes that restrict the transfer of excessive amounts of resources evolving a paternally silenced allele, and genes that aim to maximize the transfer of resources to the offspring evolving a maternally silenced allele (Forejt and Gregorová 1992; Haig 2004). The most famous example of parent-offspring conflict leading to genomic imprinting is that of the reciprocal imprinting of the *IGF2* and *IGF2R* genes (Giannoukakis et al. 1993). In humans, the maternal allele of the *IGF2* locus is methylated, leading to expression only from the paternal allele. For the *IGF2R* locus, the reverse pattern can be found. Because *IGF2* promotes embryonal growth and *IGF2R* inhibits embryonal growth, a parental conflict over the resources provided to the embryo was thought to be the driving force behind this reciprocal imprinting (Wilkins and Haig 2003).

Additionally, it has been proposed that intralocus sexual conflict can drive genomic imprinting as well (Day and Bonduriansky 2004). The idea behind this theory is that alleles that provide a sex-specific selective advantage will be more likely to be passed on to the next generation from the sex for which the advantage is present. For example, if an allele provides a male-specific advantage to find a female to mate with, this allele is more likely to be passed on to the next generation, because males having this allele will on average have more offspring. As a result of this, it is predicted to be beneficial for the offspring to express the allele received from the parent of the same sex in this case, as this allele has a higher chance of having the aforementioned selective advantage. Genomic imprinting may then emerge to ensure alleles from the opposite sex are silenced in these loci. Examples of loci where a different allele would be favored for the opposite sex would be genes that influence mate choice or courtship behavior. Therefore, this theory could explain genomic imprinting as observed in genes expressed in the brain (Davies et al. 2008; Tucci et al. 2019).

In this study, we study ASM across the genome of the livebearing fish *Poeciliopsis gracilis* (family Poeciliidae). This species supplies nutrients to offspring via yolk proteins provided before fertilization, a phenomenon known as lecithotrophy. This limits the opportunity for parent-offspring conflict, since the developing embryo does not have a direct influence on the amount of nutrients provided by the mother. However, sexual dimorphism and skewed sex ratios suggest that genes could be under the influence of sexual selection in this species (Contreras-MacBeath and Espinoza 1996). This makes *P. gracilis* an interesting case to see if ASM can be found in genes involved in sexual selection, and whether it is absent in genes involved in direct parent-offspring interaction, as predicted by the conflict theory (Haig 2000). We present a new approach for determining ASM based on the methylation-detecting abilities of the Oxford Nanopore sequencer (Simpson et al. 2017). By using reads that overlap with both a CpG site and heterozygous SNPs we can directly determine ASM, instead of inferring it indirectly from intermediate methylation levels. We sequence four individuals: two parents and two of their offspring (one male offspring, one female offspring). We hypothesize that if ASM due to sexual selection is present in this species, the female offspring will have a disproportional amount of methylation in alleles originating from the male parent, while in the male offspring the same should be the case for the female parent.

6.2 Results

6.2.1 A new approach to detect allele-specific methylation

We developed a pipeline that can detect ASM based on long-read sequencing data obtained from a single individual. For this, we use the methylation-detecting abilities of the Oxford Nanopore sequencing technology (Simpson et al. 2017). Briefly, we use read overlaps between CpG sites with nearby heterozygous SNPs to test for the hypothesis that methylated reads associate with one of the two alleles non-randomly using a hypergeometric test. Following this test for allele-specific methylation on single CpG sites, we scan the genomes for areas containing allele-specific methylation in multiple consecutive CpG sites by combining the p-values for allele-specific methylation of neighboring CpG sites using a sliding window approach. In this way, only stretches of consecutive CpG sites showing allele-specific methylation are considered as candidate regions, while single CpG sites that show a pattern of allele-specific methylation surrounded by sites that do not follow this pattern are not considered. Finally, the parent-of-origin for both haplotypes of candidate regions is found using a trio binning approach (figure 6.1).

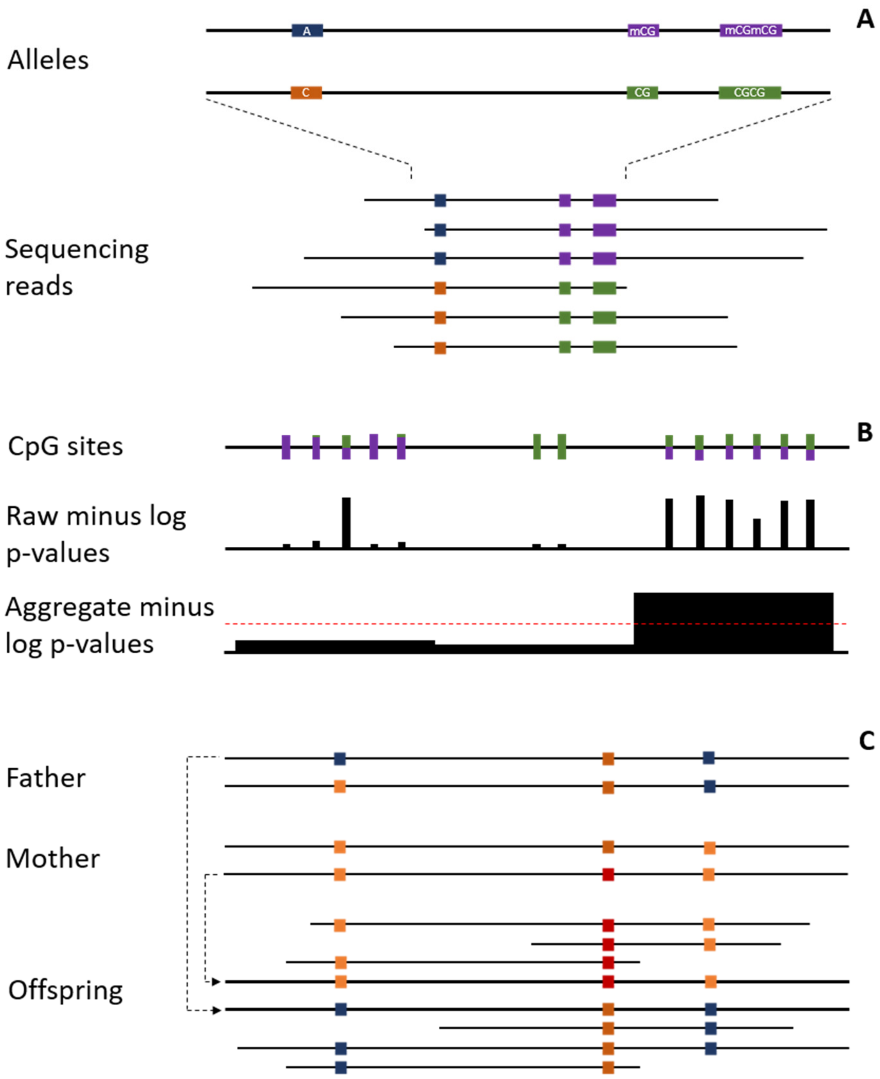

Figure 6.1 Schematic overview of our approach to detect allele-specific methylation. **A:** When long sequencing reads overlap both an allele-specific CpG site (purple/green) and a nearby heterozygous SNP in the genome (blue/orange), allele-specific methylation can be detected when both alleles of the SNP associate exclusively with only one of the two methylation states in the sequenced reads. **B:** Individual CpG sites are tested for a deviation from the expected random hypergeometric distribution (panel A). These raw p-values are aggregated using a sliding window test. **C:** Parent-of-origin of the methylated and non-methylated haplotype of the candidate regions was found by first phasing the genome of the offspring using read overlap between heterozygous SNPs (bottom half of panel), then finding the parent-of-origin of each haplotype of the phase blocks using informative SNPs (top half of panel).

6.2.2 Allele-specific methylation is widespread in *Poeciliopsis gracilis*

We applied our approach to detect allele-specific methylation in the livebearing fish *Poeciliopsis gracilis*. To do this, we sequenced four *P. gracilis* individuals to a coverage of 17-33X with Oxford Nanopore long reads (supplementary table 6.1). To find out patterns in the inheritance of ASM from parents to offspring, we sequenced two parents and two of their offspring, one male and one female offspring. We called and filtered high confidence heterozygous SNPs for each individual (supplementary table 6.1).

Using our method to detect ASM in the genome of *P. gracilis*, we found several genomic segments with strong evidence for methylation of only one of the two alleles in every individual, using a false discovery rate of 5%. In total, we found 1499 genomic segments supporting ASM for the father, 3021 for the mother, 6234 for the son, and 9949 for the daughter (supplementary table 6.2). As the parents were sequenced to a somewhat lower coverage than the offspring, the difference in genomic segments found per individual may be explained partly by a difference in statistical power to find regions of ASM. However, the results are not exactly proportional, as the son was sequenced to a higher coverage than the daughter but has fewer genomic segments with evidence for ASM.

The genomic distribution of CpG sites within regions showing ASM roughly follows the distribution of all CpG sites within the genome, although a somewhat higher than expected proportion of intergenic CpG sites shows ASM in all four individuals (figure 6.2).

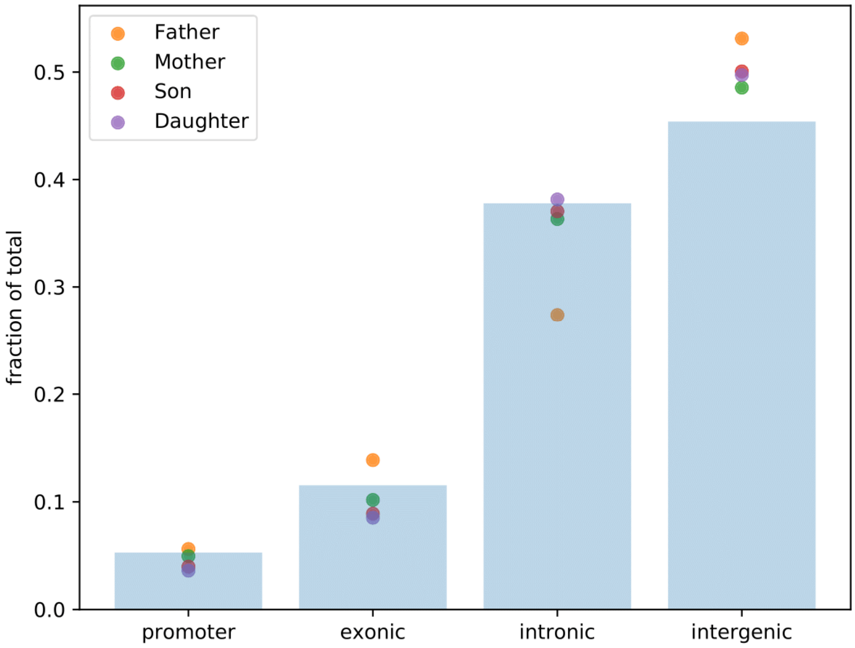

Figure 6.2 Genomic regions where CpG sites that are showing allele-specific methylation are located for the four sequenced individuals (coloured dots), compared to the genomic distribution of all CpG sites (blue bars). The promoter was defined as -1000 to +100 basepairs from the transcription start site.

6.2.3 Testing ASM hypotheses

To gain insight into how ASM inherits from parent to offspring, we phased the genomes of the offspring using read overlaps between SNPs. Then, for each phase block the parent-of-origin of each haplotype was determined using a trio binning approach (see methods). The parent-of-origin of the methylated allele of CpG sites showing ASM was then determined using read overlaps with the CpG site and the nearest heterozygous SNP.

Having parent-of-origin information of loci that exhibit ASM allows us to test the hypothesis that genomic imprinting due to intralocus sexual conflict is present in *P. gracilis*: if this were the case, in genes that are under sexual selection the methylated allele should more often than expected be inherited from the parent of the opposite sex. Indeed, when we look at the parent-of-origin of methylated alleles at sites with evidence for ASM, we do see that in the genome of the daughter more sites are methylated at the paternal allele, and in the genome of the son more sites are methylated at the maternal allele (figure 6.3A and 6.3B). However, the differences are not overwhelmingly large, suggesting that if genomic imprinting occurs in the genome of *P. gracilis*, it occurs only in a relatively small fraction of CpG sites showing ASM.

To filter out loci that were more likely involved in genomic imprinting, we made two assumptions: first, the locus should show evidence for ASM in both offspring. Second, the methylated allele in the son should be inherited from a different parent than the methylated allele in the daughter. Using these criteria, we could increase the ratio of father-to-daughter and mother-to-son methylation somewhat, relative to father-to-son and mother-to-daughter methylation. (figure 6.3C and 6.3D). We refer to these loci as crosswise methylated loci from now on.

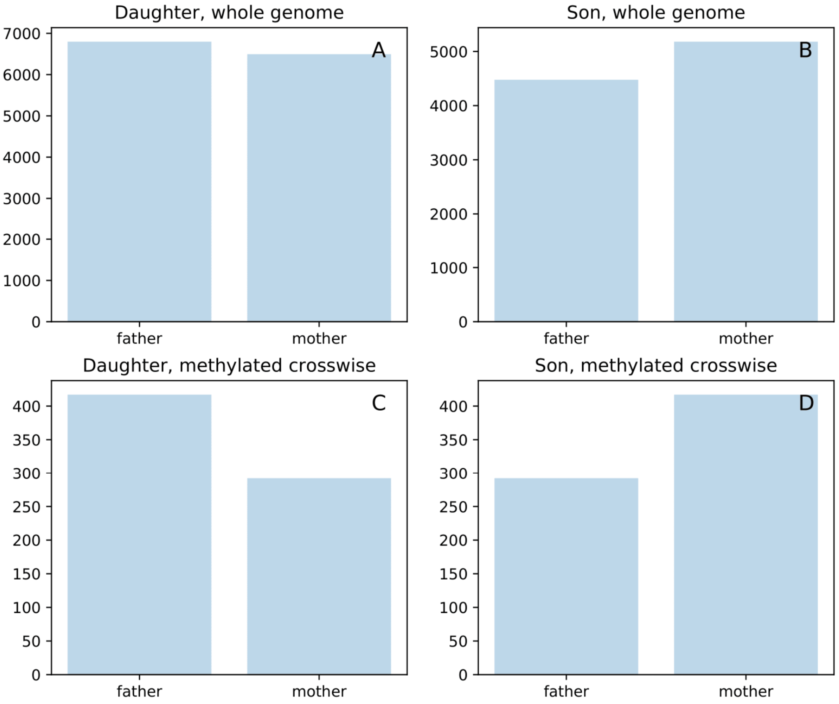

Figure 6.3 Top half: proportion of parent-of-origin of the methylated allele in all loci with significant evidence for allele-specific methylation in the genome of the daughter (figure 3A) and son (figure 3B). Bottom half: proportion of parent-of-origin of the methylated allele in all loci with significant evidence for allele-specific methylation in both son and daughter, using the additional criterium that the methylated allele from son and daughter at this locus is from

a different parent (crosswise allele-specific methylation) in the genome of the daughter (figure 3C) and son (figure 3D).

Besides genomic imprinting, we tested the hypothesis that ASM in *P. gracilis* is associated with heterozygous SNPs within CpG sites. To test this, we extracted CpG sites within regions with significant evidence for ASM and compared their heterozygosity with both all CpG sites and the genome-wide average. Indeed, CpG sites show an excess of heterozygosity in all four individuals. This difference is even more pronounced for CpG sites within regions that have significant evidence for ASM (figure 6.4A). In total, about 2.3 percent of CpG sites showing ASM in one of the two offspring contain a heterozygous SNP (figure 6.4B).

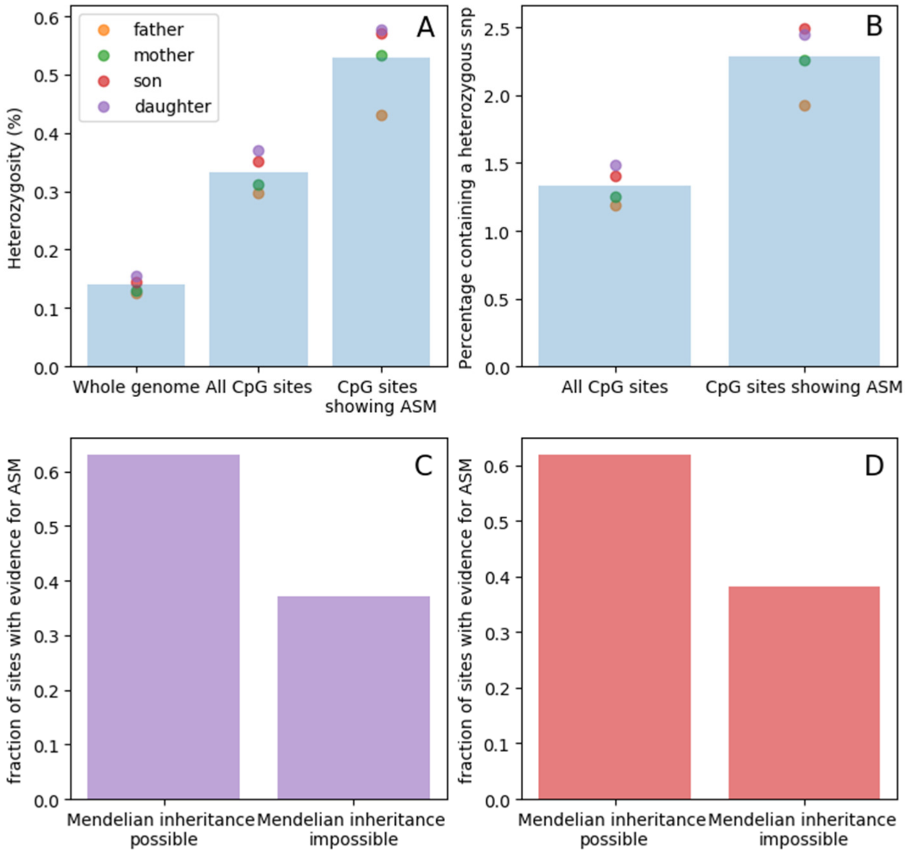

Figure 6.4 A: heterozygosity across the whole genome, across CpG sites and across CpG sites showing allele-specific methylation. Averages are in blue bars, individual data points in colored dots. **B:** Percentage of CpG sites that contain a heterozygous SNP. Averages are in blue bars, individual data points in colored dots. **C:** fraction of CpG sites with evidence for ASM following Mendelian inheritance rules in the female offspring. **D:** fraction of CpG sites with evidence for ASM following Mendelian inheritance rules in the male offspring.

Alternatively, it is possible that ASM is associated with heterozygosity not within the CpG site itself, but at a linked genomic position that regulates methylation at the CpG site. In this case, a direct link between methylation state and genomic variation is impossible to establish with our current data. However, we can still get a sense of the frequency at which this happens by reasoning that if methylation state is linked to an allele at a variable site, the methylation state at the CpG site should inherit from parents to their offspring like a "regular" allele, i.e. following Mendelian inheritance rules. When checking this assumption, we found that in both offspring, in about 35% of the cases where ASM is observed the combination of methylation states at the CpG sites in parents and offspring are in conflict with Mendelian inheritance rules (figure 6.4C and 6.4D). This indicates that although genetic variation can partially explain ASM in *P. gracilis*, for a substantial portion of CpG sites the methylation state is not linked to genetic variation but is regulated by other mechanisms, i.e. epigenetically regulated.

6.2.4 Functional enrichment analysis

To find genes that may be under the influence of genomic imprinting, we extracted every gene from the *P. gracilis* genome that is either overlapping or lies directly downstream of our candidate regions (see methods for selection criteria). The candidate regions were defined as regions where (1) both offspring show ASM at the same locus, and (2) the methylated allele in both offspring is inherited from the opposite parent, and (3) no heterozygous SNPs were present inside CpG sites within the candidate regions. This led to the identification of 182 genes that are potentially under the influence of genomic imprinting in the genome of *P. gracilis* (supplementary table 6.3).

Functional enrichment analysis of this gene set revealed that a greater than expected proportion of the candidate gene set is located in the brain, specifically in the post-synaptic membrane (supplementary table 6.4).

6.2.5 Clustering of loci with evidence for allele-specific methylation

To gain more insight in the genomic distribution of ASM on a larger scale, we plotted the density of ASM occurrence over the genome of *P. gracilis* (see figure 6.5 for the ten longest scaffolds). We found that ASM is relatively widespread in *P. gracilis*, with most scaffolds having occurrences of ASM across the whole scaffolds, and specific clusters of ASM could not be observed. However, when looking at our imprinting candidates (crosswise methylation), we observe a much sparser distribution, with occurrences of crosswise methylation mostly being limited to a few larger clusters (figure 6.5, bottom panel).

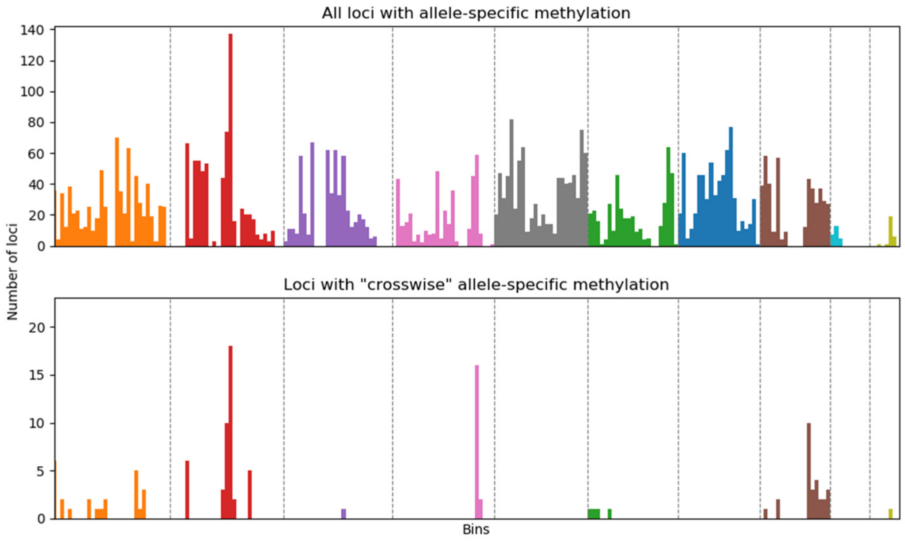

Figure 6.5 For the ten longest scaffolds of the P. gracilis assembly, for 500 kb bins, number of regions with evidence for allele-specific methylation (top panel), as well as the number of regions with evidence for allele-specific methylation, and the methylated allele for the son originates from the opposite parent as the methylated allele for the daughter (crosswise methylation, bottom panel). Only parts of the scaffolds that could be phased are shown in this picture.

Within these clusters, evidence for ASM is very strong in CpG sites around all adjacent genes. An example of this the promoter of the *baiap2* gene, which is involved in brain insulin metabolism and dendrite formation (Oda et al. 1999; Kang et al. 2016). In our dataset, evidence for ASM at the promoter of this gene is very high for three out of four individuals (figure 6.6), with only the father showing a completely unmethylated promoter region.

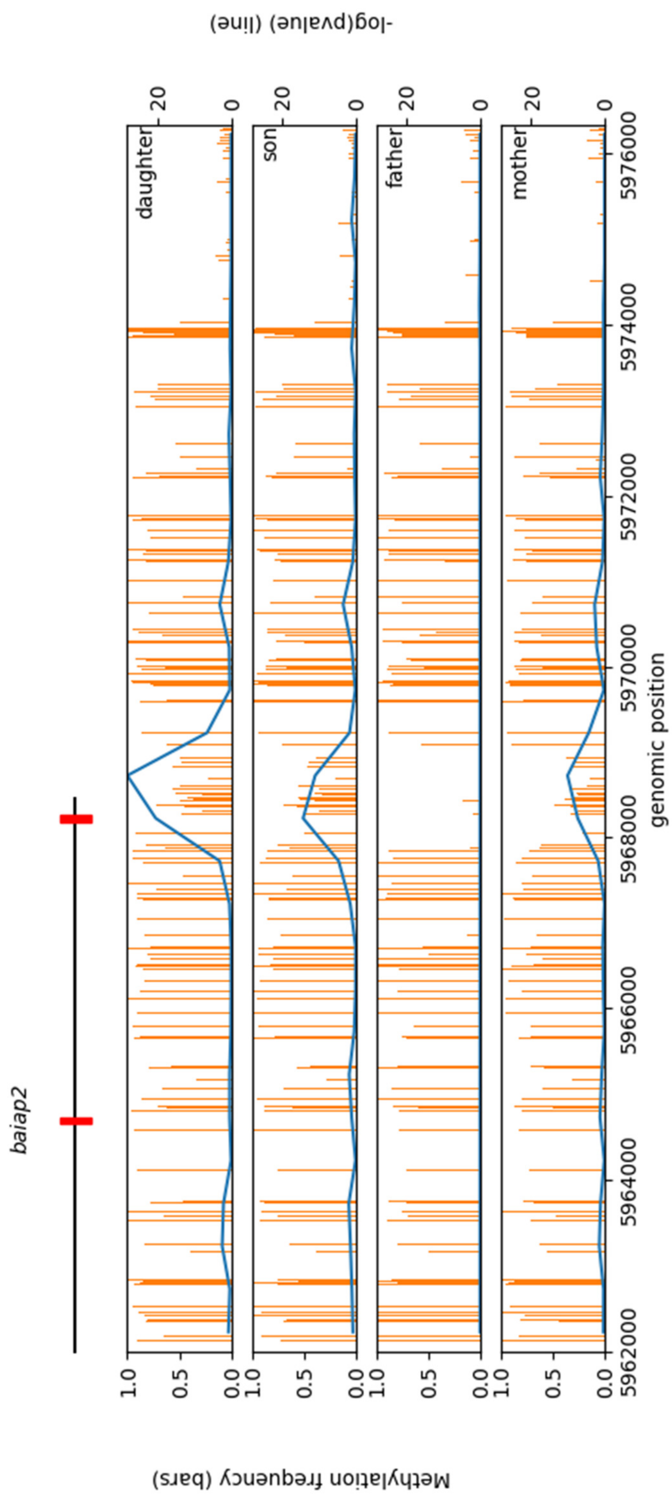

Figure 6.6 Methylation frequency across CpG sites (orange bars) and the minus-log p-value for the test for allele-specific methylation (blue line) at the promoter of the *baiap2* gene in the genome of the four sequenced individuals.

6.3 Discussion

In this study, we show ASM in the genome of four related *P. gracilis* individuals using a newly developed approach based on methylation calls from nanopore sequencing. With this method, ASM can be observed directly using reads that overlap both a CpG site and a heterozygous SNP. This allows for a more direct way to observe ASM compared to classical bisulfite sequencing, as sequencing only a single individual is now sufficient to detect ASM. Additionally, sequencing the CpG site and the heterozygous SNP on the same physical molecule ensures that the different methylation states are linked to the different alleles. This is more reliable than using bisulfite sequencing to detect ASM, as mixing cells with different methylation states at the same locus can introduce false positives when using bisulfite sequencing to detect ASM as the reads are too short to directly infer haplotypes.

Our results show that ASM is widespread in the genome of *P. gracilis*, as we found that 824 to 3442 coding genes show evidence for ASM in nearby regions, depending on the individual (supplementary table 6.2). Likely, the true number of genes with nearby ASM is even higher, as with our method we could only test for ASM if the CpG site was within reasonable distance of a heterozygous SNP. Because the used population of *P. gracilis* is highly inbred, around 40% of the genome consists of runs of homozygosity that could not be tested for ASM (supplementary figure 6.1). According to our tests, both parents showed substantially less genes with evidence for ASM around them than the offspring: 824 and 1513 genes for the father and mother, versus 2986 and 3442 genes for the son and daughter (supplementary table 6.2). Part of this may be due to differences in sequencing depth, as the parents were sequenced to a somewhat lower sequencing depth than their offspring. This suggests that genomic segments having weaker evidence for ASM, for instance because of tissue-specific ASM, can only be found with a high sequencing depth: if not all sequenced tissues contain ASM, only a fraction of all reads will contain support for ASM, and a higher sequencing depth is necessary to detect these instances of ASM. However, many of the clusters showing ASM in the offspring show either complete methylation or complete absence of methylation for either one or both parents across CpG islands that show intermediate levels of methylation in the offspring (see figure 6 for example), a result that cannot be explained by differences in sequencing depth. Age-specific methylation differences may be the reason behind these observations, as this processes has been observed in humans (Day et al. 2013; McCarthy et al. 2014).

Based on parent-of-origin tests of CpG sites showing ASM we can see that within the genome of the son, methylation originates slightly more frequently from the maternal genome than from the paternal genome, and vice versa for the genome of the daughter. These differences become larger when we select for sites for which the methylated allele is from the opposite parent for the son and daughter, something we call crosswise methylation. Genes around these sites are highly enriched for their localization at the synapse, something that is somewhat remarkable given that we sequenced a mixture of fin and muscle tissue for all individuals. The most straightforward explanation for this observation is that the methylation at these sites was already present within the gametes from which the individuals grew, so it is now present in all tissues, as is known to occur in mammals (Bourc'his et al. 2001; Hata et al. 2002). However, the ASM will likely only have a physiological effect in tissues where the surrounding genes are expressed, which would in this case be the brain.

Currently, there are three known reasons for allele-specific methylation: sex chromosome inactivation, heterozygous SNP-induced allele-specific methylation and genomic imprinting (Kaslow and Migeon 1987; Li et al. 1993b; Shoemaker et al. 2010). Sex chromosome inactivation seems unlikely as an explanation for the majority of our results because (i) the clusters of allele-specific methylation are scattered across almost all scaffolds of the assembly, and (ii) fish from the family Poeciliidae have much smaller genomic differences between male and female individuals than observed in mammals, as they do not have “classical” sex chromosomes (Haaf and Schmid 1984; Kottler et al. 2020).

Heterozygosity can partly explain our observations, as we find approximately a threefold increase in heterozygosity between CpG sites that show ASM and the genome-wide average. At the heterozygous site, methylation of the CpG site is disrupted by the mutation at one allele, leading to ASM. We find a heterozygous site in approximately 2.3 percent of our CpG sites inside regions showing ASM, but the proportion of regions showing ASM because of heterozygosity may be higher, as it has been shown that ASM in humans can also show linkage with sequence variants outside the CpG site (Hutchinson et al. 2014). However, given the fact that for about 40 percent of CpG sites with evidence for ASM, Mendelian inheritance of methylation state is impossible, variation in the genomic sequence cannot be the sole explanation for ASM in *P. gracilis*.

The third known mechanism for ASM is genomic imprinting. In mammals, genomic imprinting is often attributed to a conflict over resources allocated from mother to offspring (Monk 2015). This makes our study system interesting: although *P. gracilis* is a livebearing fish, it does not have a placenta, and all nutrients from which the embryo grows are supplied before fertilization in the form of egg yolk. This limits the influence the embryo can exert on the mother's nutrient supply, suggesting that a direct conflict over nutrient allocation cannot be a driving factor towards the observed ASM. Besides a conflict over resources, it has been predicted that intralocus sexual conflict, i.e. sex-specific selection pressure on a certain locus, can also drive the evolution of genomic imprinting (Day and Bonduriansky 2004). Given the extensive occurrence of sexual selection in poeciliid fish (Pollux et al. 2014), it is likely that some genes will have different selective pressures for male and female individuals in this family. This is particularly true for genes that influence reproductive success differently in each sex, such as genes involved in behavioral traits.

Although definitive evidence for genomic imprinting at specific loci would require the sequencing of multiple additional parent-offspring pairs, we found a number of results that suggest the presence of genomic imprinting in this species. First, our list of candidate genes contains several genes for which different selection pressures are predicted between males and females: genes involved in brain functioning, consistent with sex-specific selection on behavioral traits. Second, CpG sites within regions with evidence for ASM are more often than expected methylated in a "crosswise" fashion, e.g. the methylated allele in the son is inherited from the mother, and the methylated allele in the daughter is inherited from the father. This pattern has been predicted to occur at loci under the influence of genomic imprinting due to intralocus sexual conflict (Day and Bonduriansky 2004). Third, heterozygosity seems to only partially explain the ASM that we find, given the deviation from Mendelian inheritance of methylation state in a substantial portion of our candidate CpG sites. These findings suggest that genomic imprinting due to intralocus sexual conflict contributes to ASM in *P. gracilis*, and invites further study to investigate this hypothesis.

6.4 Methods

6.4.1 Sequencing of four *Poeciliopsis gracilis* individuals

One male and one female *Poeciliopsis gracilis* individual were kept in a single tank. After offspring was born, the two parents were sacrificed using a lethal dose of MS-222. The offspring were grown for three months in the same tank that housed the parents, until they reached a size large enough for sex determination. Of the offspring, one male and one female offspring were sacrificed using the same method as for the parents. For each individual, DNA was extracted from a mixture of fin and muscle tissue using the Qiagen genomic-tip 100/G kit.

After extraction, short DNA fragments (<10kb) were depleted from the samples using the Circulomics Short Read Eliminator kit. Oxford Nanopore libraries were prepared using the Rapid Barcoding Kit (SQK-RBK004) and the four samples were sequenced on a single PromethION flowcell, and basecalled using Guppy 4.3.4 at High Accuracy Mode.

6.4.2 Read mapping and variant calling

Sequencing reads of the four sequenced individuals were mapped to the *P. gracilis* genome assembly (available at the European Nucleotide Archive under accession number PRJEB46512) using minimap2 version 2.17 using the map-ont setting for mapping nanopore reads (Li 2018). Candidate SNPs were called using nanopolish version 0.11.3 (Simpson 2018), with the nanopolish variants command on all four alignment files separately, using default settings. As nanopore base calling is less accurate at indels, only substitutions were called. All sites in the genome for which a SNP was called in at least one individual were then re-called for all four individuals simultaneously using freebayes v1.2.0 with the setting --use-best-n-alleles 4 (Garrison and Marth 2012). The output VCF of this program was then filtered for calling quality > 30 and average read depth between 10 and 80 to prevent false positive SNP calls due to structural variants.

6.4.3 Methylation calling

To find both methylated and unmethylated CpG sites, methylation calling was performed using nanopolish version 0.11.3 (Simpson 2018) on the genome alignments using default settings. An index was created linking sequence files with their raw signal files using nanopolish index. After that, methylation was called using nanopolish call-methylation, using a likelihood ratio threshold of 2.

6.4.4 Finding allele-specific methylation

From the methylation calls, candidate CpG sites for allele-specific methylation (ASM) were selected based on a methylation frequency between 0.35 and 0.65. Because Oxford Nanopore reads are multiple kilobases long, several reads can be found that overlap both the ambiguously methylated site and the closest heterozygous site. If the ambiguous methylation on this site is due to ASM, we would expect all reads that show the reference allele at the heterozygous site to have a different methylation status at the methylation site than the reads that show the alternative allele at the heterozygous site. If the ambiguous methylation on this site is not due to ASM, we would expect no correlation between methylation status on the methylation site and allele at the heterozygous site. We tested for deviation from expected distribution by using a hypergeometric test, which gives a probability that our observations occur by chance, assuming no ASM. We calculated this probability for each CpG site in our genome for which we could find at least six reads supporting methylation status at a CpG site and the associated allele at the closest heterozygous site.

If genomic regions, rather than individual sites, display ASM, it is expected that multiple sites near each other show consistent separation between methylation state and alleles. To find these regions, we aggregated the p-values among all called CpG sites in a 500 base window using Fisher’s p-value aggregation method, sliding 100 bases for each test. In this way, a single CpG site with a low p-value surrounded by CpG sites that seem to follow the hypergeometric distribution is not considered as a candidate region, while stretches of multiple sites showing skew from the distribution are considered.

6.4.5 Finding parent-of-origin of offspring haplotypes

To find the parent-of-origin of both alleles in SNPs and in CpG sites showing ASM, the SNPs of the offspring individuals were first phased using read overlaps using whatsHap 1.1 using default settings (Martin et al. 2016), yielding phase blocks in which alleles could be reliably categorized into haplotypes. Because the *P. gracilis* individuals used were highly inbred, only about 60 percent of the genome could be phased because of a lack of heterozygous SNPs in some areas. Therefore, we only continued with genomic regions that could be phased. For each phase block, parent genotypes at offspring SNP locations were used to deduce which haplotype was passed on from which parent to the offspring, if the SNP was informative.

6.4.6 Gene Ontology enrichment tests

Each gene for which a candidate CpG site was either within the coding region of the gene, within the intronic part of the gene, or within 50kb upstream of the gene, was considered. The predicted amino acid sequence for each gene was extracted based on the gene annotation. The String Database (version 11.0) was used to find human orthologs of these sequences, and subsequently to perform GO enrichment tests on this orthologous gene set.

6.5 Declarations

6.5.1 Ethics approval

All animal-derived biological materials used this study were obtained according to local ethical regulations.

6.5.2 Consent for publication

Not applicable

6.5.3 Availability of data and materials

All sequencing reads generated for this study, as well as the *P. gracilis* genome assembly are available at the European Nucleotide Archive under Bioproject number PRJEB46512.

All supplementary material referenced to in this thesis are available on the Zenodo database (10.5281/zenodo.5647272).

6.5.4 Competing interests

The authors declare no competing interests.

6.5.5 funding

This work was supported by a WIAS research grand from Wageningen Institute of Animal Sciences, awarded to HJM and BJAP, and a VIDI grant from the Netherlands Organisation for Scientific Research (864.14.008) awarded to BJAP.

6.5.6 Author contributions

TE performed breeding and rearing of the *P. gracilis* individuals. HvK and RN performed DNA extraction, library preparation and sequencing of the *P. gracilis* individuals. All bioinformatic analyses and writing of the initial manuscript were performed by HvK, supervised by HJM and BJAP. RN, TE, OM, MAMG, BJAP and HJM critically reviewed the manuscript for further improvements.

6.5.7 Acknowledgements

Not applicable

7

General discussion

7.1 Introduction

Unravelling the genomic basis of complex trait evolution is one of the greatest challenges in evolutionary biology. Due to the recent decrease in sequencing costs and the accompanying surge in available genome assemblies, there is now an excellent opportunity to study this phenomenon (Wetterstrand 2021). In this thesis, I have investigated the genomic basis of the evolution of one such complex trait, the placenta in the livebearing fish family Poeciliidae. Here, I discuss my findings.

For this thesis, I have sequenced and assembled the genomes of 18 poeciliid fish species using a variety of sequencing and assembly techniques. In section **7.2** I discuss my experiences in the sequencing and assembly of new genomes, comparing the results of different techniques. Also, I discuss the trade-off between the cost of a sequencing technique and the resulting assembly quality: which assembly type is "good enough" for which analysis? After assembling the genomes, I have compared the genomes of placental and non-placental species. In section **7.3** I discuss some of the comparative genomics methods that I used. In section **7.4** I discuss the biological interpretation of my results: after four years of research, what can we say about the genomics of placenta evolution in livebearing fish? In this section, I will go into some of the limitations of this study as well. Finally, in section **7.5** I discuss the future of this research topic. Here I propose a strategy that aims to answer some of the questions this thesis has raised.

7.2 Sequencing and genome assembly

For this thesis, the genomes of 18 poeciliid species have been assembled, using a variety of sequencing techniques (table 7.1). Thirteen of these genomes were sequenced using Illumina short-reads, which was at the time of sequencing the cheapest method by a wide margin. This technique is nowadays mainly used for resequencing of individuals of a species for which a genome assembly is already available, as assembling a genome of short reads only yields a highly fragmented assembly. Indeed, genomes assembled in this thesis with short reads only show a low scaffold N50 (6.4-65.1 Kb) compared to the assemblies based on third-generation sequencing techniques (1.8-21.6 Mb). Still, around 80% of all single-copy genes could be retrieved in one piece in these genomes (table 7.1), indicating that most scaffold breaks are located in non-genic, likely repetitive, regions. When reference quality genomes are available of closely related species, orthologs of single-copy genes can be retrieved with high accuracy using alignment-based methods (chapter 4). Sequencing many genomes using this cheap technique instead of sequencing few genomes with a more expensive, high-quality sequencing technique has, for chapter 4 in particular, increased statistical power for comparative methods considerably. Therefore, despite advances in the sequencing field, it may still be worth considering short read assemblies when interested only in single-copy gene evolution. However, for the analysis of structural variants these assemblies are not suitable, as they are too fragmented.

Table 7.4. Assembly statistics for all genomes assembled in this thesis

Species	Sequencing technique	Read length N50 (bp)	Scaffold N50	BUSCO complete %
Poeciliopsis retropinna	PacBio	16963	21.6 Mb	97.5
Poeciliopsis turrubarensis	10X genomics	150	4.2 Mb	95.5
Poeciliopsis turneri	Oxford Nanopore	12052	1.8 Mb	97.2
Poeciliopsis gracilis	Oxford Nanopore	6792	4.6 Mb	97.7
Phalloptychus januarius	Oxford Nanopore	8728	13.0 Mb	97.4
Alfaro cultratus	Illumina	150	57.4 Kb	86.3
Brachyrhaphis roseni	Illumina	150	41.6 Kb	82.3
Brachyrhaphis terrabensis	Illumina	150	6.4 Kb	54.9
Girardinus metallicus	Illumina	150	41.8 Kb	82.3
Heterandria formosa	Illumina	150	27.0 Kb	76.1
Poecilia bifurca	Illumina	150	21.4 Kb	73.9
Poecilia gilii	Illumina	150	25.8 Kb	80.3
Poecilia picta	Illumina	150	22.1 Kb	74.9
Poecilia wingei	Illumina	150	16.7 Kb	73.1
Poeciliopsis infans	Illumina	150	59.6 Kb	88.5
Poeciliopsis paucimaculata	Illumina	150	65.1 Kb	90.1
Poeciliopsis presidionis	Illumina	150	38.7 Kb	84.4
Poeciliopsis prolifica	Illumina	150	38.6 Kb	83.1

Five genomes were assembled using third-generation sequencing techniques that should, in theory, produce reference-quality assemblies. The genome of *Poeciliopsis retropinna* was assembled using PacBio long-reads (Rhoads and Au 2015), which has yielded a genome assembly of excellent contiguity, better than the genomes assembled with 10X genomics or Oxford Nanopore data. In fact, many of the *P. retropinna* chromosomes were assembled in one contig, without the help of additional techniques that link assembled contigs to physical chromosomes (chapter 3). Additionally, almost all universal single-copy genes were present in this genome. However, PacBio sequencing was, at least at the time of sequencing, more expensive than 10X genomics or Oxford Nanopore sequencing. The genome of *Poeciliopsis turrubarensis* was assembled using 10X genomics linked reads. This is not a "true" long read sequencing technique. Instead, short reads originating from the same DNA molecule are given the same barcode before sequencing, which can be used to infer genomic proximity during the assembly procedure. This procedure yielded an assembly with a much greater contiguity than traditional short-read assemblies. However, gene completeness was lacking slightly when compared to the PacBio and Oxford Nanopore assemblies. Finally, the genomes of *Poeciliopsis turneri*, *Poeciliopsis gracilis* and *Phalloptychus januarius* were assembled using Oxford Nanopore long reads. These genomes, although not quite as contiguous as the PacBio assembly, are still contiguous enough for structural variant analysis, as shown in chapter 5, and show excellent gene completeness. An additional advantage is that this is currently the cheapest of the three techniques.

Besides genomic sequencing data, with Oxford Nanopore sequencing it is possible to detect cytosine methylation (chapter 6). DNA methylation is important for gene regulation, and incorporation of this data type in a comparative framework will likely yield additional insights. Therefore, sequencing genomic DNA using Oxford Nanopore sequencing has the potential to be more cost-efficient than sequencing techniques where detecting cytosine methylation is not possible in comparative projects, as this will yield two data types with a single analysis. With PacBio sequencing, direct DNA methylation detection is also possible (Flusberg et al. 2010). However, with Illumina-based sequencing techniques such as 10X genomics, this is not possible as these sequencing techniques are PCR-based, and the PCR multiplication procedure erases the DNA methylation signature that is present in the original genomic DNA.

7.3 Methods in comparative genomics

Throughout this thesis, I have used several comparative genomics methods to compare the genomes of placental with non-placental species. First, I used codon- and amino acid-based methods to compare protein-coding sequence evolution specifically. Secondly, I used evolutionary constraint based methods to compare patterns of genomics evolution outside protein-coding regions. Finally, I used a whole genome alignment based method to find structural variants. Here, I will discuss my experiences with the different methods.

To detect gene evolution deviating from the expected degree of purifying selection that usually occurs on protein-coding genes, I have used two separate methods throughout my thesis: d_N/d_S-based analysis as implemented in the PAML package (Yang 2007), and evolutionary rate analysis, for which I wrote my own implementation. For the first two research chapters (chapters 2 and 3), I focused on d_N/d_S-based methods, as for these methods only a single case genome is required. In the case of the Poeciliidae, that would be one genome of a placental species, with several genomes of non-placental species to create a background distribution. As several genomes of non-placental poeciliid species were already available but none of placental poeciliids, the analysis could already be performed immediately when the first placental genomes were assembled. Although d_N/d_S analysis did yield several interesting candidate genes (chapter 2 and 3), results remain suggestive because only a single placental origin was investigated at the same time. Even when multiple origins were included as foreground branches, the test as implemented in PAML does not explicitly test for a greater degree of genomic convergence than expected by chance. In chapter 4, I tried to assess this problem by performing the same analysis on a control group of random branches in the phylogeny, for which no genomic convergence would be expected. Here, no excess of low p-values was observed in placental species, compared to the control group. Additionally, it is known that d_N/d_S based methods lose power when comparing very closely related species (Gharib and Robinson-Rechavi 2013). In the poeciliid phylogeny, several but not all placental species have closely related non-placental relatives, leading to short terminal branches in the phylogeny. Therefore, it is challenging to make meaningful conclusions about genomic convergence using the results of this analysis: do species-specific differences in the results of this analysis reflect a difference in selective pressure, or just a difference in phylogenetic topology? For these reasons, d_N/d_S analysis seems not suitable as a method to assess genomic convergence associated with phenotypic convergence.

A method specifically designed to assess genomic convergence linked to phenotypic variation is the evolutionary rate analysis. First used to assess genomic changes associated with adaptations to underwater life in marine mammals (Chikina et al. 2016), this method originally compares the evolutionary rate of the amino acid sequence of a protein-coding gene between species in two groups: one group that has the trait of interest, and one that does not. This is different from d_N/d_S analysis because synonymous mutations in the coding sequence are ignored, as they do not lead to a change in the amino acid sequence. Evolutionary rates are normalized for gene- and species- specific evolutionary rates, so that deviations from the expected evolutionary rates are compared instead of the raw mutation counts. In the Poeciliidae, it is hard to categorize all species in a group that has a placenta and a group that does not, as a spectrum of placental complexities is observed rather than species that have a fully complex placenta and species that do not have a placenta at all (Reznick et al. 2002; Pollux et al. 2009). Therefore, instead of dividing the branches in the phylogeny in two groups, I modified this method to test for a correlation between the relative evolutionary rate of a gene in all branches of the phylogeny and the Matrotrophy Index (MI) of the species the branches are leading to. I recommend this adapted method for any study that tries to correlate evolutionary rate with a trait that can be easily represented by a continuous variable.

To find consistent genomic changes in non-coding genomic regions, I employed a method to detect evolutionary constraint in two multi-genome alignments of poeciliid fish as implemented in the PhastCons program (Hubisz et al. 2010). This program compares the evolutionary rate of all sites in a multi genome alignment with that of neutrally evolving sequences, such as ancient repetitive elements or fourfold degenerate sites. A decrease in evolutionary rate compared to the baseline would indicate evolutionary constraint, suggesting functional relevance of the site. Aggregates of nearby genomic sites with high evidence for evolutionary constraint are clustered into so-called conserved elements: stretches of genomic sequence that are predicted to be functional. This method has shown to be accurate in determining functional non-coding elements in the genomes of model species, such as humans (King et al. 2005). However, the length and frequency of predicted conserved elements are highly influenced by the expected-length and target-coverage parameters in the PhastCons program that have to be submitted by the user, and are therefore somewhat subjective when used for species of which no reasonable predictions for these parameters exist. In chapter 4, I did not analyze the PhastCons output directly, but instead analyzed the differences between two individual PhastCons analyses: one based on a multi genome alignment with only highly matrotrophic species, and one based on a multi genome alignment with only lecithotrophic species. In addition to enabling a more robust comparative genomics study, this is a good way to measure the consistency of the PhastCons output: performing two identical analyses on very closely related species should yield very similar results. Indeed, more than 99% of the conserved elements were called for both multi genome alignments (chapter 4), indicating the robustness of conserved element analysis for this dataset.

Finally, I used two different methods to find structural variants: a read mapping based method in chapter 2, and a whole genome alignment based method that I developed in chapter 3, and extended for multi-genome comparisons in chapter 5. Read mapping based analysis is a widely used method to find structural variants, and many tools are available (Abyzov et al. 2011; Rausch et al. 2012; Layer et al. 2014; Ye et al. 2018). Generally, this method is used to find structural variants within populations of a single species, by resequencing several individuals and running the analysis for all individuals in parallel. In chapter 2, I aimed to use the method of finding structural variants through read mappings cross-species, by mapping reads from the placental *Heterandria formosa* to the genome of the non-placental *Poecilia reticulata*, and then calling structural variants using a read depth- and split read-based method. Although this method yielded some compelling candidates, conservative filtering was necessary to filter out false positives due to mis-mapping of reads, as the genomic difference between these species was greater than the difference between within-species individuals that this method is generally used for. Additionally, several genomic regions present in both *P. reticulata* and *H. formosa* were diverged to the point where larger-scale synteny could still be identified, but individual reads would no longer map cross-species, leading to false positive deletion calls. This made it necessary to omit all deletions completely from the final result, as they could not be called accurately. Because of these reasons, I would not recommend read mapping based methods to call structural variants across species.

In chapter 3 and 5, I used a different approach to find structural variants, based on the alignment of whole assembled genomes, not individual reads. This approach is not new, and has been applied before on great ape genomes (Kronenberg et al. 2018). For this approach, pairwise whole genome alignments are scanned for genomic areas that deviate from the expected 1:1 syntenic relationship. Based on the type of deviation a duplication, deletion or rearrangement is called. When more than two genomes are present, generally pairwise alignments to a single reference, usually an outgroup, are made to find structural variants for each target species sequentially. For chapter 5, I developed an implementation of this method that is based on the reference-free whole genome aligner Cactus (Paten et al. 2011). This implementation has two advantages over previous methods. First, using a reference-free alignment means that structural variants can be validated against multiple outgroups simultaneously, which increases the robustness of the results and decreases reference bias. Second, I included a method to select for structural variants that occur in multiple species at the same genomic location simultaneously, which allows for testing the hypothesis that certain structural variants are consistently associated with placenta evolution. Although this method is more reliable than short read mapping based methods for the identification of structural variants between species, certain weaknesses still have to be kept in mind. Specifically, this method is prone to false positives due to assembly errors, especially when the compared species have been sequenced using a different sequencing technique. Therefore, using high quality assemblies to find structural variants is essential. If using high quality assemblies, setting a minimum contig length for the identification of structural variants, as well as a minimum distance from the beginning or the end of a contig will filter out the majority of false positives.

7.4 Biological conclusions and limitations

7.4.1 Main findings of this thesis

I believe this thesis has illuminated several aspects of the genomics of placenta evolution in poeciliid fish. First, results presented in this thesis show that the evolution of the poecilid placenta is accompanied by consistent accelerated evolution of the protein sequence of metabolic and structural genes, as well as a change in the regulatory sequences around developmental genes. With that, the work done on this thesis provides the first evidence for family-wide genomic convergence associated with placenta evolution in the Poeciliidae. These findings are not surprising, as the combination of coding changes in metabolic and structural genes and regulatory changes around developmental genes as a basis for complex trait evolution has been predicted several times before, and has also been observed in other studies (Prud'homme et al. 2007; Carroll 2008; Chikina et al. 2016; Partha et al. 2017). It is clear that, in vertebrates, many genes involved in early development have so many functions that a change in amino acid sequence would almost always be deleterious due to pleiotropic effects. Regulatory changes can potentially circumvent these pleiotropic effects while still affecting the phenotype in early development.

Second, there is evidence that the evolution of a rudimentary placenta accelerates its own further evolution, given that most genes associated with placenta evolution maintain or even further accelerate their evolutionary rate even after a rudimentary placenta has formed (chapter 4). This is consistent with predictions about placenta evolution being accelerated by the emergence of a parent-offspring conflict (Crespi and Semeniuk 2004; Pollux et al. 2009), and explains why a placenta has been gained nine times in the family Poeciliidae, but has never been lost.

Third, gene duplications do not seem to play a consistent role in the process of placenta evolution, as duplicated genomic segments were not shared between placental species more than expected by chance. Gene duplications are another way to, in theory, circumvent deleterious pleiotropic effects when changing the amino acid sequence of a protein with many functions. However, in the case of placenta evolution in poeciliid fish this does not seem to be happening. We did find an association of placenta evolution with the deletions of certain types of genes, such as genes involved in the calcineurin-NFAT signaling cascade or the BMP signaling pathway. These deletions are hard to interpret biologically, as genomic deletions are classically not associated with the evolution of new complex functions. Additionally, because of the teleost-specific whole genome duplication, most of the deleted genes have a second copy that is still intact. Currently, there is nowhere near enough knowledge about how these paralogs function to make reliable predictions about the effect of such deletions. If I were to speculate, subfunctionalization of paralogs would be the most likely reason for paralogous genes to co-exist after a whole genome duplication (Lynch and Force 2000). A deletion of one of these copies after subfunctionalization would disturb the associated molecular pathway, but only in cell types in which this copy is expressed. Possibly, these deletions may therefore be a representation of the pruning of certain pathways in tissue types relevant to placental evolution, such as ovary or liver tissue.

Fourth, the dynamics of allele-specific DNA methylation in *Poeciliopsis gracilis* suggest the presence of genomic imprinting in the genome of this species. This result seems somewhat counterintuitive at first, as genomic imprinting is classically associated with parent-offspring conflict in a placental species (Zeh and Zeh 2001; Crespi and Semeniuk 2004). However, as discussed in chapter 6, genomic imprinting has also been predicted to occur as a result of sex-specific selection (Day and Bonduriansky 2004), something that occurs throughout the Poeciliidae and is not exclusive to placental species. These results show that poeciliid genomes have the potential for complex methylation dynamics. In placental species, for which a parent-offspring conflict is predicted, methylation may have an effect on the expression of genes involved in placentation. Therefore, as a possible future research avenue, studying allele-specific methylation in placental poeciliids may give important insights on parent-offspring conflict in these species. The results from chapter 6 show that this is feasible, and, with Oxford Nanopore being able to call methylation from a "regular" sequencing run, very cost-efficient.

7.4.2 Questions raised by this thesis

A few questions remain without a definitive answer, although the results from this thesis at least allow for a more educated guess than before. First, although our results show genomic convergence in genes with certain functions associated with placenta evolution, it is hard to pinpoint exactly the mechanistic changes occurring in molecular pathways that cause the evolution of a placenta. I believe this is the single largest question that remains unanswered after the completion of this thesis. Comparative genomics methods are very powerful for showing that certain genomic patterns can be observed in association with a trait of interest, but to prove that a single distinct change in a molecular pathway causes a certain phenotypic response, more concrete evidence is necessary. Speculating, it would seem logical that the change in tissue structure observed in extensive matrotrophic poeciliids is achieved by changing the regulation of developmental genes that induce epithelial folding so that they are expressed at the follicular epithelium. The principle of epithelial folding is essential for the development of the vertebrate body (Zartman and Shvartsman 2010), therefore all molecular tools to provide such a tissue change should be present in the genome, requiring only a change in the time and place of expression for placental tissue to form. Several methods could facilitate the testing of this hypothesis, with RNA-Seq, ChiP-Seq and ATAC-Seq being a few examples of techniques that can illuminate the wirings of a regulatory network and help with the functional annotation of a genome. Integrating these data types into a comparative -omics framework will be the next challenge for future researchers working on this subject, and I propose one way of approaching this problem in section 7.5.

Second, we cannot be sure of the exact functions of candidate genes that are identified using comparative genomics methods in a non-model organism. The currently used functional annotation methods rely on sequence similarity to experimentally curated proteins for a prediction of function. However, differences in functional domain composition are not usually taken into account when assigning a gene name or putative function, especially in complex gene families (see figure 7.1 for example). Modern annotation pipelines like MAKER or BRAKER (Campbell et al.

2014; Hoff et al. 2019) do perform scanning for functional domains, but the results are just added to the output as is, and are not taken into account for the assignment of a predicted gene name and function. This problem is magnified in fish genomes, because of the teleost-specific genome duplication. Because this genome duplication is fish-specific, some paralogous proteins with visibly different domain compositions will be assigned the same name and function, just because they point back to the same human protein based on sequence similarity. As a result of this, candidate gene sets of comparative genomics analyses have to be analyzed by looking at broad overrepresentations of certain biological functions, to prevent single mis-annotations of specific molecular functions from influencing the interpretation of results. Curating more Zebrafish proteins experimentally will help with this problem somewhat, and a functional domain-aware annotation pipeline, one that also incorporates domain predictions into gene name and function prediction, would be beneficial to dealing with this problem as well.

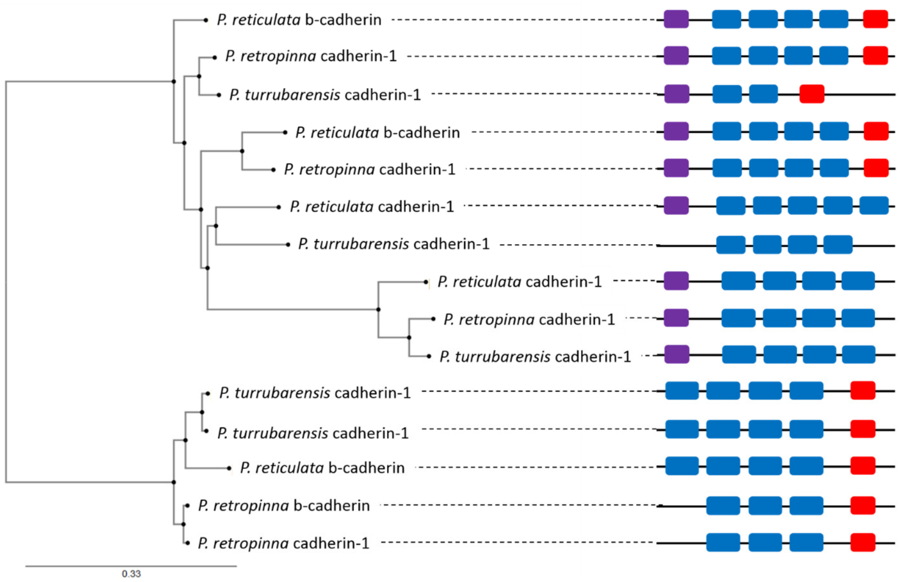
P. reticulata b-cadherin
P. retropinna cadherin-1
P. turrubarensis cadherin-1
P. reticulata b-cadherin
P. retropinna cadherin-1
P. reticulata cadherin-1
P. turrubarensis cadherin-1
P. reticulata cadherin-1
P. retropinna cadherin-1
P. turrubarensis cadherin-1
P. turrubarensis cadherin-1
P. turrubarensis cadherin-1
P. reticulata b-cadherin
P. retropinna b-cadherin
P. retropinna cadherin-1
0.33

Figure 7.4. Phylogenetic relationship and domain architecture of genes annotated as cadherin-1 or b-cadherin in three poeciliid fish (*P. retropinna*, *P. turrubarensis* and *P. reticulata*). As a result of gene annotation via sequence similarity, assigned genes names do not correlate well with phylogenetic relationships or domain architecture.

Third, it is unlikely that the results presented in this thesis reflect all changes that occur when a placenta evolves. For instance, I assumed phenotypic convergence that is fully reflected in genomic convergence for many of the analyses. However, from morphological studies it is known that placental structures from different evolutionary origins are actually not exactly the same, but show some differences in morphology and rate of embryo weight gain (Guernsey et al. 2020; Zandonà et al. 2021)(Bart Pollux, personal communication). It is therefore likely that some genomic changes related to placenta evolution will be unique for each origin. Possibly, I have shown some of these unique changes using the positive selection analyses used throughout the thesis (chapter 2, 3 and 4), as these can be applied to a single phylogenetic branch. Ironically, genomic convergence is necessary to statistically show that these changes are related to placenta evolution, and for changes that are unique for a certain origin this genomic convergence will be absent.

7.5 Future directions: a network-based approach to pinpoint specific regulatory differences between placental and non-placental poeciliids

After showing that the evolution of the placenta is associated with both genic and regulatory changes in the genomes of placental species, the next step towards a greater understanding of complex trait evolution is pinpointing exactly how regulatory pathways change in placental species. This will provide understanding of the mechanism of placenta evolution, rather than a list of genes that is associated with placenta evolution.

To achieve this goal, integrating new data types into one functional comparative framework will be essential. Here, I propose one approach to infer the regulatory changes involved in placenta evolution by modeling the protein-DNA interactions of candidate genes in a biological network, after which it is possible to compare the networks of placental species with those of non-placental species. The approach can be summarized in four steps: (1) identifying proteins containing DNA binding domains using functional annotation, (2) predicting DNA binding domain – DNA sequence motif pairs for these proteins with the help of functional data, (3) constructing a directional protein-DNA network based on the genomic location of sequence motifs for each species separately, and (4) using network alignment to find regulatory differences between placental and non-placental species.

7.5.1 identifying proteins containing DNA binding domains using functional annotation

As discussed in section 7.4, the identification of protein function based on sequence similarity to proteins in model species is error-prone. However, the identification of functional domains in a protein sequence is much more reliable due to the fact that functional domains are generally more conserved than a protein sequence as a whole. Specifically, DNA binding domains can be identified with close to 90% accuracy (Mishra et al. 2019). This domain-centric approach allows us to find a set of proteins with regulatory potential. Note that while DNA binding domains can be identified computationally with high accuracy inside a predicted coding sequence, the coding sequence itself cannot be predicted with high accuracy without functional data. Therefore, for a high-quality prediction of gene models, gene expression data such as RNA-seq is essential. Long-read sequencing techniques can aid in providing high-quality gene model predictions, as with these techniques RNA or cDNA can be sequenced with the whole transcript being on a single read (Sessegolo et al. 2019; Workman et al. 2019), which increases the accuracy of both intron and isoform detection substantially.

7.5.2 Predicting DNA binding domain – DNA sequence motif pairs

After DNA binding proteins have been identified, the next step to model a regulatory network is to predict and curate DNA sequence motifs to which the proteins bind. For many DNA binding domains, such information is already available in databases such as ENPD (Tak Leung et al. 2019) or uniPROBE (Newburger and Bulyk 2009). These databases summarize available information on functional protein domains and the DNA motifs to which they bind, and include experimentally validated information on several fish species, including the poeciliid fish *Poecilia reticulata*. Putative domain-sequence pairs can be filtered based on available evidence, and extra evidence can be provided with both functional data generation and comparative genomics analyses. For important candidate genes, ChIP-seq experiments can be performed to show that proteins indeed bind to certain DNA motifs, and results can also be cross-checked with a conserved elements analysis to show that candidate motifs are recurring in different fish species.

7.5.3 Constructing a regulatory network model

Once domain-DNA motif binding pairs have been predicted, a regulatory network can be constructed by drawing directed edges between the genes having DNA binding domains and genes directly downstream of the DNA motifs predicted to be binding sites for these domains. The challenge here is to separate genomic locations with protein-binding DNA motifs that actually bind a protein from those that do not, as protein-binding DNA motifs are more plentiful than actual enhancers. Therefore, the selection of connections has to be refined by integrating evidence from functional or comparative genomics data.

Genomic regions containing active enhancers can be identified using ATAC-seq experiments, something that has been successfully done in several studies (Thibodeau et al. 2018; Bozek et al. 2019). ATAC-seq data may be combined with sequence conservation data from whole genome alignments, to test whether the candidate site is conserved in multiple species. Additionally, predicted regulatory connections may be validated by performing an RNA-seq co-expression experiment, where we can test whether the expression of one gene really affects the expression of a gene predicted to be downstream of this gene in a regulatory cascade. These kinds of co-expression data are widely used for the construction biological networks, but are also known for generating many non-physical interactions, just because co-expression does not necessarily mean physical interaction (Yu et al. 2013). However, if used to validate a potential physical interaction of a protein domain and a DNA motif by looking at the co-expression of a nearby gene, co-expression data can refine a model of a regulatory network significantly.

Approaches to construct gene regulatory networks from functional data are plenty, and excellent reviews are available (Delgado and Gómez-Vela 2019; Wang et al. 2021). Comparing different methods extensively is beyond the scope of this discussion, but two particularly relevant points from the aforementioned reviews are that (1) methods that incorporate multiple sources of functional data generally perform better, and (2) regardless of the method used, large-scale (whole genome) regulatory network reconstructions contain high levels of noise even when modeling the simplest of organisms. By contrast, local regulatory networks can be modeled with relatively high accuracy, especially if key regulatory candidates are validated experimentally. For instance, local regulatory networks have, with the help of additional functional data, been modeled successfully for the regulation of blood cell development (Moignard et al. 2015; Pina et al. 2015).

Considering these factors, I propose to model the protein-DNA interactions and construct a local regulatory network around candidate genes identified with comparative genomics analyses, such as available from the results of this thesis. Performing this analysis for placental and non-placental poeciliid species separately, the goal is to find consistent differences in the resulting networks. For this, network alignment methods can be used.

7.5.4 Network alignment

After networks are constructed for each species separately, the goal is to look for consistent network differences between placental and non-placental species, which would indicate regulatory differences that coincide with placenta evolution. Aligning the regulatory networks of different poeciliid species is one way to find these consistent differences. Network alignment is a well-established method, for which numerous tools have been developed (Guzzi and Milenković 2018; Ma and Liao 2020). The principle of network alignment is to cluster nodes in the networks of different species together based on their interaction patterns of their neighboring communities, so that genes with orthologous function can be identified based on their similarities in interaction patterns. As with sequence alignment, local and global alignment methods exist, which aim to maximize local subgraph topological similarity and global network topological similarity, respectively. As the aim is to model specific developmental pathway, local network alignment seems preferrable here.

Using aligned networks, potential regulatory differences between placental and non-placental poeciliids can be extracted easily by looking at consistent differences in network topology between placental and non-placental species. An edge between two nodes that is present in all placental species but absent in all non-placental species is a candidate for regulatory change associated with placenta evolution. Because in the proposed approach the edges represent a regulatory interaction between two genes, a consistent difference in the network between placental and non-placental species will be easier to interpret biologically than when looking at gene evolution only.

7.6 Concluding remarks

The advent of third-generation sequencing techniques has brought the field of comparative genomics a newfound significance. In this thesis, I show the power of the comparative genomics approach while investigating the evolution of the placental in the livebearing fish family Poeciliidae. I show that placenta evolution in this family is associated with consistent genomic change, pointing towards regulatory change in developmental regions, as well as changes in the protein-coding regions of metabolic and structural genes.

As prices of different -omics techniques continue to decrease, it is inevitable that future studies will incorporate different kinds of functional data in combination with comparative genomics analyses to further our understanding of complex trait evolution. When these different data types are integrated properly into a comparative framework, I believe that our understanding of complex trait evolution will grow even further.

References

Ababneh M, Troedsson M. 2013. Ovarian steroid regulation of endometrial phospholipase A2 isoforms in horses. *Reproduction in Domestic Animals* **48**: 311-316.

Abyzov A, Urban AE, Snyder M, Gerstein M. 2011. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* **21**: 974-984.

Adiconis X, Borges-Rivera D, Satija R, DeLuca DS, Busby MA, Berlin AM, Sivachenko A, Thompson DA, Wysoker A, Fennell T. 2013. Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nat Methods* **10**: 623.

Ahn C, Yang H, Lee D, An B-s, Jeung E-B. 2015. Placental claudin expression and its regulation by endogenous sex steroid hormones. *Steroids* **100**: 44-51.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**: 403-410.

Anderson JP, Gleason CA, Foley RC, Thrall PH, Burdon JB, Singh KB. 2010. Plants versus pathogens: an evolutionary arms race. *Funct Plant Biol* **37**: 499-512.

Aplin J, Jones C, Harris L. 2009. Adhesion molecules in human trophoblast–a review. I. Villous trophoblast. *Placenta* **30**: 293-298.

Arancibia-Garavilla Y, Toledo F, Casanello P, Sobrevia L. 2003. Nitric oxide synthesis requires activity of the cationic and neutral amino acid transport system y+ L in human umbilical vein endothelium. *Experimental physiology* **88**: 699-710.

Armstrong J, Hickey G, Diekhans M, Fiddes IT, Novak AM, Deran A, Fang Q, Xie D, Feng S, Stiller J. 2020. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* **587**: 246-251.

Ball MP, Li JB, Gao Y, Lee J-H, LeProust EM, Park I-H, Xie B, Daley GQ, Church GM. 2009. Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat Biotechnol* **27**: 361-368.

Banet A, Reznick D. 2008. Do placental species abort offspring? Testing an assumption of the Trexler–DeAngelis model. *Funct Ecol* **22**: 323-331.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**: 455-477.

Barlow DP, Bartolomei MS. 2014. Genomic imprinting in mammals. *Cold Spring Harbor perspectives in biology* **6**: a018382.

Basolo AL. 1990. Female preference predates the evolution of the sword in swordtail fish. *Science* **250**: 808-810.

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL. 2004. The Pfam protein families database. *Nucleic Acids Res* **32**: D138-D141.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol*: 289-300.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**: 573-580.

Bianconi ME, Dunning LT, Moreno-Villena JJ, Osborne CP, Christin P-A. 2018. Gene duplication and dosage effects during the early emergence of C4 photosynthesis in the grass genus Alloteropsis. *J Exp Bot* **69**: 1967-1980.

Bisazza A. 1993. Male competition, female mate choice and sexual size dimorphism in poeciliid fishes. *Marine & Freshwater Behaviour & Phy* **23**: 257-286.

Blackburn DG. 2015. Evolution of vertebrate viviparity and specializations for fetal nutrition: a quantitative and qualitative analysis. *J Morphol* **276**: 961-990.

Boeckmann B, Bairoch A, Apweiler R, Blatter M-C, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'donovan C, Phan I. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* **31**: 365-370.

Boetzer M, Pirovano W. 2014. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinformatics* **15**: 211.

Boller T, He SY. 2009. Innate immunity in plants: an arms race between pattern recognition receptors in plants and effectors in microbial pathogens. *Science* **324**: 742-744.

Bourc'his D, Xu G-L, Lin C-S, Bollman B, Bestor TH. 2001. Dnmt3L and the establishment of maternal genomic imprints. *Science* **294**: 2536-2539.

Bozek M, Cortini R, Storti AE, Unnerstall U, Gaul U, Gompel N. 2019. ATAC-seq reveals regional differences in enhancer accessibility during the establishment of spatial coordinates in the Drosophila blastoderm. *Genome Res* **29**: 771-783.

Brown HN, Gale BH, Johnson JB, Belk MC. 2018. Testes mass in the livebearing fish *Brachyrhaphis rhabdophora* (Poeciliidae) varies hypoallometrically with body size but not between predation environments. *Ecology and Evolution* **8**: 11656-11662.

Bshary R, Wickler W, Fricke H. 2002. Fish cognition: a primate's eye view. *Animal cognition* **5**: 1-13.

Buhl W, Eisenlohr L, Preuss I, Gehring U. 1995. A novel phospholipase A2 from human placenta. *Biochem J* **311**: 147.

Campbell MS, Holt C, Moore B, Yandell M. 2014. Genome annotation and curation using MAKER and MAKER-P. *Current protocols in bioinformatics* **48**: 4.11. 11-14.11. 39.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**: 1972-1973.

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

Cavallotti I, De Luca L, D Aponte A, De Falco M, Acanfora F, Visciano M, Gualdiero L, De Luca B, Baldi A, De Luca A. 2001. Expression of the retinoblastoma-related p107 and Rb2/p130 genes in human placenta: An imunohistochemical study. *Histology and histopathology* **16**: 1057-1060.

Chen Z, Cheng C-HC, Zhang J, Cao L, Chen L, Zhou L, Jin Y, Ye H, Deng C, Dai Z. 2008. Transcriptomic and genomic evolution under constant cold in Antarctic notothenioid fish. *Proceedings of the National Academy of Sciences* **105**: 12944-12949.

Chikina M, Robinson JD, Clark NL. 2016. Hundreds of genes experienced convergent shifts in selective pressure in marine mammals. *Mol Biol Evol* **33**: 2182-2192.

Cimino MC. 1973. Karyotypes and erythrocyte sizes of some diploid and triploid fishes of the genus *Poeciliopsis*. *J Fish Res Board Can* **30**: 1736-1737.

Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA, Tanenbaum DM, Civello D, Lu F, Murphy B. 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* **302**: 1960-1963.

Clevers H. 2006. Wnt/β-catenin signaling in development and disease. *Cell* **127**: 469-480.

Contreras-MacBeath T, Espinoza HR. 1996. Some aspects of the reproductive strategy of Poeciliopsis gracilis (Osteichthyes: Poeciliidae) in the Cuautla River, Morelos, Mexico. *J Freshwat Ecol* **11**: 327-338.

Crabtree GR, Schreiber SL. 2009. SnapShot: Ca2+-calcineurin-NFAT signaling. *Cell* **138**: 210.

Crespi B, Semeniuk C. 2004. Parent-offspring conflict in the evolution of vertebrate reproductive mode. *Am Nat* **163**: 635-653.

Crick F. 1970. Central dogma of molecular biology. *Nature* **227**: 561-563.

Cross J, Baczyk D, Dobric N, Hemberger M, Hughes M, Simmons D, Yamamoto H. 2003. Genes, development and evolution of the placenta. *Placenta* **24**: 123-130.

Cross JC, Nakano H, Natale DR, Simmons DG, Watson ED. 2006. Branching morphogenesis during development of placental villi. *Differentiation* **74**: 393-401.

Dabney A, Storey JD, Warnes G. 2010. qvalue: Q-value estimation for false discovery rate control. *R package version* **1**.

Daley WP, Matsumoto K, Doyle AD, Wang S, DuChez BJ, Holmbeck K, Yamada KM. 2017. Btbd7 is essential for region-specific epithelial cell dynamics and branching morphogenesis in vivo. *Development*: dev. 146894.

Darwin C. 1859. On the origin of the species by natural selection.

Daub J, Moretti S, Davydov II, Excoffier L, Robinson-Rechavi M. 2017. Detection of pathways affected by positive selection in primate lineages ancestral to humans. *Mol Biol Evol* **34**: 1391-1402.

Davies W, Isles AR, Humby T, Wilkinson LS. 2008. What are imprinted genes doing in the brain? In *Genomic imprinting*, pp. 62-70. Springer.

Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comp Biol* **6**: e1001025.

Day K, Waite LL, Thalacker-Mercer A, West A, Bamman MM, Brooks JD, Myers RM, Absher D. 2013. Differential DNA methylation with age displays both common and dynamic features across human tissues that are influenced by CpG landscape. *Genome Biol* **14**: 1-19.

Day T, Bonduriansky R. 2004. Intralocus sexual conflict can drive the evolution of genomic imprinting. *Genetics* **167**: 1537-1546.

Delgado FM, Gómez-Vela F. 2019. Computational methods for gene regulatory networks reconstruction and analysis: a review. *Artif Intell Med* **95**: 133-145.

DeMarais A, Oldis D. 2005. Matrotrophic transfer of fluorescent microspheres in poeciliid fishes. *Copeia* **2005**: 632-636.

Deng H-X, Shi Y, Yang Y, Ahmeti KB, Miller N, Huang C, Cheng L, Zhai H, Deng S, Nuytemans K. 2016. Identification of TMEM230 mutations in familial Parkinson's disease. *Nat Genet* **48**: 733.

Deng L, Feng J, Broaddus R. 2010. The novel estrogen-induced gene EIG121 regulates autophagy and promotes cell survival under stress. *Cell death & disease* **1**: e32.

Dial KP. 2003. Wing-assisted incline running and the evolution of flight. *Science* **299**: 402-404.

Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. 2009. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**: 48.

Ekström P, Meissl H. 2003. Evolution of photosensory pineal organs in new light: the fate of neuroendocrine photoreceptors. *Philosophical Transactions of the Royal Society B: Biological Sciences* **358**: 1679-1700.

Farrell JA, Wang Y, Riesenfeld SJ, Shekhar K, Regev A, Schier AF. 2018. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* **360**.

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

Fleuren M, Quicazan-Rubio EM, van Leeuwen JL, Pollux BJ. 2018. Why do placentas evolve? Evidence for a morphological advantage during pregnancy in live-bearing fish. *PloS one* **13**: e0195976.

Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW. 2010. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods* **7**: 461-465.

Foote AD, Liu Y, Thomas GW, Vinař T, Alföldi J, Deng J, Dugan S, van Elk CE, Hunter ME, Joshi V. 2015. Convergent evolution of the genomes of marine mammals. *Nat Genet* **47**: 272.

Forejt Ji, Gregorová Sa. 1992. Genetic analysis of genomic imprinting: an Imprintor-1 gene controls inactivation of the paternal copy of the mouse Tme locus. *Cell* **70**: 443-450.

Frey N, Frank D, Lippl S, Kuhn C, Kögler H, Barrientos T, Rohr C, Will R, Müller OJ, Weiler H. 2008. Calsarcin-2 deficiency increases exercise capacity in mice through calcineurin/NFAT activation. *The Journal of clinical investigation* **118**: 3598-3608.

Furness AI, Pollux BJ, Meredith RW, Springer MS, Reznick DN. 2019. How conflict shapes evolution in poeciliid fishes. *Nature communications* **10**: 1-12.

Gaetje R, Holtrich U, Engels K, Kissler S, Rody A, Karn T, Kaufmann M. 2008. Differential expression of claudins in human endometrium and endometriosis. *Gynecological Endocrinology* **24**: 442-449.

Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. 2009. Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25**: i54-i62.

Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:12073907*.

Gartler SM, Riggs AD. 1983. Mammalian X-chromosome inactivation. *Annu Rev Genet* **17**: 155-190.

Gaunt S. 2002. Conservation in the Hox code during morphological evolution. *Int J Dev Biol* **38**: 549-552.

Gemayel R, Cho J, Boeynaems S, Verstrepen KJ. 2012. Beyond junk-variable tandem repeats as facilitators of rapid evolution of regulatory and coding sequences. *Genes* **3**: 461-480.

Ghalambor CK, Reznick DN, Walker JA. 2004. Constraints on adaptive evolution: the functional trade-off between reproduction and fast-start swimming performance in the Trinidadian guppy (Poecilia reticulata). *The American Naturalist* **164**: 38-50.

Gharib WH, Robinson-Rechavi M. 2013. The branch-site test of positive selection is surprisingly robust but lacks power under synonymous substitution saturation and variation in GC. *Mol Biol Evol* **30**: 1675-1686.

Giannoukakis N, Deal C, Paquette J, Goodyer CG, Polychronakos C. 1993. Parental genomic imprinting of the human IGF2 gene. *Nat Genet* **4**: 98-101.

Glasauer SM, Neuhauss SC. 2014. Whole-genome duplication in teleost fishes and its evolutionary consequences. *Mol Genet Genomics* **289**: 1045-1060.

Gobert M, Lafaille JJ. 2012. Maternal-fetal immune tolerance, block by block. *Cell* **150**: 7-9.

Graef IA, Chen F, Chen L, Kuo A, Crabtree GR. 2001. Signals transduced by Ca2+/calcineurin and NFATc3/c4 pattern the developing vasculature. *Cell* **105**: 863-875.

Griffith OW, Wagner GP. 2017. The placenta as a model for understanding the origin and evolution of vertebrate organs. *Nature ecology & evolution* **1**: 0072.

Grove BD, Wourms JP. 1991. The follicular placenta of the viviparous fish, Heterandria formosa. I. Ultrastructure and development of the embryonic absorptive surface. *J Morphol* **209**: 265-284.

Grove BD, Wourms JP. 1994. Follicular placenta of the viviparous fish, Heterandria formosa: II. Ultrastructure and development of the follicular epithelium. *J Morphol* **220**: 167-184.

Guernsey MW, van Kruistum H, Reznick DN, Pollux BJ, Baker JC. 2020. Molecular signatures of placentation and secretion uncovered in Poeciliopsis maternal follicles. *Mol Biol Evol* **37**: 2679-2690.

Guzzi PH, Milenković T. 2018. Survey of local and global biological network alignment: the need to reconcile the two sides of the same coin. *Briefings in bioinformatics* **19**: 472-481.

Gwack Y, Sharma S, Nardone J, Tanasa B, Iuga A, Srikanth S, Okamura H, Bolton D, Feske S, Hogan PG. 2006. A genome-wide Drosophila RNAi screen identifies DYRK-family kinases as regulators of NFAT. *Nature* **441**: 646-650.

Haaf T, Schmid M. 1984. An early stage of ZW/ZZ sex chromosome differentiation in *Poecilia sphenops* var. melanistica (Poeciliidae, Cyprinodontiformes). *Chromosoma* **89**: 37-41.

Hagmayer A, Furness AI, Reznick DN, Dekker ML, Pollux BJ. 2020. Predation risk shapes the degree of placentation in natural populations of live-bearing fish. *Ecol Lett* **23**: 831-840.

Hahn C, Bachmann L, Chevreux B. 2013. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucleic Acids Res* **41**: e129-e129.

Haig D. 1993. Genetic conflicts in human pregnancy. *Q Rev Biol* **68**: 495-532.

Haig D. 2000. The kinship theory of genomic imprinting. *Annu Rev Ecol Syst* **31**: 9-32.

Haig D. 2004. Genomic imprinting and kinship: how good is the evidence? *Annu Rev Genet* **38**: 553-585.

Hardison RC. 2003. Comparative genomics. *PLoS Biol* **1**: e58.

Hata K, Okano M, Lei H, Li E. 2002. Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice.

Heard E, Clerc P, Avner P. 1997. X-chromosome inactivation in mammals. *Annu Rev Genet* **31**: 571-610.

Hickey G, Paten B, Earl D, Zerbino D, Haussler D. 2013. HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics* **29**: 1341-1342.

Hinz B, Gabbiani G. 2003. Cell-matrix and cell-cell contacts of myofibroblasts: role in connective tissue remodeling. *Thrombosis and haemostasis* **89**: 993-1002.

Hirano S, Yan Q, Suzuki ST. 1999. Expression of a novel protocadherin, OL-protocadherin, in a subset of functional systems of the developing mouse brain. *J Neurosci* **19**: 995-1005.

Hirase S, Ozaki H, Iwasaki W. 2014. Parallel selection on gene copy number variations through evolution of three-spined stickleback genomes. *BMC Genomics* **15**: 735.

Hittinger CT, Carroll SB. 2007. Gene duplication and the adaptive evolution of a classic genetic switch. *Nature* **449**: 677-681.

Hoegg S, Meyer A. 2005. Hox clusters as models for vertebrate genome evolution. *Trends Genet* **21**: 421-424.

Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. 2019. Whole-Genome Annotation with BRAKER. In *Gene Prediction*, pp. 65-95. Springer.

Hoffberg SL, Troendle NJ, Glenn TC, Mahmud O, Louha S, Chalopin D, Bennetzen JL, Mauricio R. 2018. A High-Quality Reference Genome for the Invasive Mosquitofish *Gambusia affinis* Using a Chicago Library. *G3: Genes, Genomes, Genetics*: g3. 200101.202018.

Hojayev B, Rothermel BA, Gillette TG, Hill JA. 2012. FHL2 binds calcineurin and represses pathological cardiac growth. *Mol Cell Biol* **32**: 4025-4034.

Holland PW, Garcia-Fernàndez J, Williams NA, Sidow A. 1994. Gene duplications and the origins of vertebrate development. *Development* **1994**: 125-133.

Hou Z, Romero R, Uddin M, Than NG, Wildman DE. 2009. Adaptive history of single copy genes highly expressed in the term human placenta. *Genomics* **93**: 33-41.

Hubisz MJ, Pollard KS, Siepel A. 2010. PHAST and RPHAST: phylogenetic analysis with space/time models. *Briefings in bioinformatics* **12**: 41-51.

Hughes AL, Green JA, Garbayo JM, Roberts RM. 2000. Adaptive diversification within a large family of recently duplicated, placentally expressed genes. *Proceedings of the National Academy of Sciences* **97**: 3319-3323.

Hutchinson JN, Raj T, Fagerness J, Stahl E, Viloria FT, Gimelbrant A, Seddon J, Daly M, Chess A, Plenge R. 2014. Allele-specific methylation occurs at genetic variants associated with complex disease. *PloS one* **9**: e98464.

Illingworth RS, Bird AP. 2009. CpG islands–'a rough guide'. *FEBS Lett* **583**: 1713-1720.

Jaime Zúñiga-Vega J, N Reznick D, B Johnson J. 2007. Habitat predicts reproductive superfetation and body shape in the livebearing fish Poeciliopsis turrubarensis. *Oikos* **116**: 995-1005.

Jia R-Z, Zhang X, Hu P, Liu X-M, Hua X-D, Wang X, Ding H-J. 2012. Screening for differential methylation status in human placenta in preeclampsia using a CpG island plus promoter microarray. *Int J Mol Med* **30**: 133-141.

Jollie WP, Jollie LG. 1964. The fine structure of the ovarian follicle of the ovoviviparous poeciliid fish, Lebistes reticulatus. II. Formation of follicular pseudoplacenta. *J Morphol* **114**: 503-525.

Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**: 1236-1240.

Jue NK, Foley RJ, Reznick DN, O'Neill RJ, O'Neill MJ. 2018. Tissue-Specific Transcriptome for Poeciliopsis prolifica Reveals Evidence for Genetic Adaptation Related to the Evolution of a Placental Fish. *G3: Genes, Genomes, Genetics*: g3. 200270.202018.

Kang J, Park H, Kim E. 2016. IRSp53/BAIAP2 in dendritic spine development, NMDA receptor regulation, and psychiatric disorders. *Neuropharmacology* **100**: 27-39.

Kaslow DC, Migeon BR. 1987. DNA methylation stabilizes X chromosome inactivation in eutherians but not in marsupials: evidence for multistep maintenance of mammalian X dosage compensation. *Proc Natl Acad Sci* **84**: 6210-6214.

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

Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. 2003. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proceedings of the National Academy of Sciences* **100**: 11484-11489.

Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. 2011. Adaptive seeds tame genomic sequence comparison. *Genome Res* **21**: 487-493.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**: 907-915.

Kim JC, Mirkin SM. 2013. The balancing act of DNA repeat expansions. *Curr Opin Genet Dev* **23**: 280-288.

Kim MJ, Deng H-X, Wong YC, Siddique T, Krainc D. 2017. The Parkinson's disease-linked protein TMEM230 is required for Rab8a-mediated secretory vesicle trafficking and retromer trafficking. *Hum Mol Genet* **26**: 729-741.

King DC, Taylor J, Elnitski L, Chiaromonte F, Miller W, Hardison RC. 2005. Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. *Genome Res* **15**: 1051-1060.

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

Kokot M, Długosz M, Deorowicz S. 2017. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics* **33**: 2759-2761.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* **37**: 540.

Konczal M, Przesmycka KJ, Mohammed RS, Phillips KP, Camara F, Chmielewski S, Hahn C, Guigo R, Cable J, Radwan J. 2020. Gene duplications, divergence and recombination shape adaptive evolution of the fish ectoparasite Gyrodactylus bullatarudis. *Mol Ecol* **29**: 1494-1507.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**: 722-736.

Koski MK, Haapalainen AM, Hiltunen JK, Glumoff T. 2004. A two-domain structure of one subunit explains unique features of eukaryotic hydratase 2. *J Biol Chem* **279**: 24666-24672.

Kottler VA, Feron R, Nanda I, Klopp C, Du K, Kneitz S, Helmprobst F, Lamatsch DK, Lopez-Roques C, Lluch J. 2020. Independent Origin of XY and ZW Sex Determination Mechanisms in Mosquitofish Sister Species. *Genetics* **214**: 193-209.

Krishnan J, Athar F, Rani TS, Mishra RK. 2017. Simple sequence repeats showing 'length preference'have regulatory functions in humans. *Gene* **628**: 156-161.

Kronenberg ZN, Fiddes IT, Gordon D, Murali S, Cantsilieris S, Meyerson OS, Underwood JG, Nelson BJ, Chaisson MJ, Dougherty ML. 2018. High-resolution comparative analysis of great ape genomes. *Science* **360**.

Kumar RP, Krishnan J, Singh NP, Singh L, Mishra RK. 2013. GATA simple sequence repeats function as enhancer blocker boundaries. *Nature communications* **4**: 1844.

Kumar RP, Senthilkumar R, Singh V, Mishra RK. 2010. Repeat performance: how do genome packaging and regulation depend on simple sequence repeats? *Bioessays* **32**: 165-174.

Kunkel TA. 2004. DNA replication fidelity. *J Biol Chem* **279**: 16895-16898.

Künstner A, Hoffmann M, Fraser BA, Kottler VA, Sharma E, Weigel D, Dreyer C. 2016. The genome of the Trinidadian guppy, *Poecilia reticulata*, and variation in the Guanapo population. *PloS one* **11**: e0169087.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. *Genome biology* **5**: R12.

Kwan L, Fris M, Rodd FH, Rowe L, Tuhela L, Panhuis TM. 2015. An examination of the variation in maternal placentae across the genus Poeciliopsis (Poeciliidae). *J Morphol* **276**: 707-720.

Lamb TD, Collin SP, Pugh EN. 2007. Evolution of the vertebrate eye: opsins, photoreceptors, retina and eye cup. *Nature Reviews Neuroscience* **8**: 960-976.

Lawton BR, Sevigny L, Obergfell C, Reznick D, O'Neill RJ, O'Neill MJ. 2005. Allelic expression of IGF2 in live-bearing, matrotrophic fishes. *Dev Genes Evol* **215**: 207-212.

Layer RM, Chiang C, Quinlan AR, Hall IM. 2014. LUMPY: a probabilistic framework for structural variant discovery. *Genome biology* **15**: R84.

Lechner M, Findeiß S, Steiner L, Marz M, Stadler PF, Prohaska SJ. 2011. Proteinortho: detection of (co-) orthologs in large-scale analysis. *BMC Bioinformatics* **12**: 124.

Lee C, Wevrick R, Fisher R, Ferguson-Smith M, Lin C. 1997. Human centromeric DNAs. *Hum Genet* **100**: 291-304.

Levin M, Anavy L, Cole AG, Winter E, Mostov N, Khair S, Senderovich N, Kovalev E, Silver DH, Feder M. 2016. The mid-developmental transition and the evolution of animal body plans. *Nature* **531**: 637.

Li E, Beard C, Jaenisch R. 1993a. Role for DNA methylation in genomic imprinting. *Nature* **366**: 362.

Li E, Beard C, Jaenisch R. 1993b. Role for DNA methylation in genomic imprinting. *Nature* **366**: 362-365.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094-3100.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**: 1754-1760.

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

Liu S-F, Lu G-X, Liu G, Xing X-W, Li L-Y, Wang Z. 2004. Cloning of a full-length cDNA of human testis-specific spermatogenic cell apoptosis inhibitor TSARG2 as a candidate oncogene. *Biochem Biophys Res Commun* **319**: 32-40.

Löytynoja A. 2014. Phylogeny-aware alignment with PRANK. In *Multiple sequence alignment methods*, pp. 155-170. Springer.

Löytynoja A, Goldman N. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* **320**: 1632-1635.

Lu P, Takai K, Weaver VM, Werb Z. 2011. Extracellular matrix degradation and remodeling in development and disease. *Cold Spring Harbor perspectives in biology* **3**: a005058.

Lynch M. 2002. Gene duplication and evolution. *Science* **297**: 945-947.

Lynch M, Force A. 2000. The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154**: 459-473.

Ma C-Y, Liao C-S. 2020. A review of protein–protein interaction network alignment: From pathway comparison to global alignment. *Computational and Structural Biotechnology Journal* **18**: 2647.

Madsen EM, Lindegaard ML, Andersen CB, Damm P, Nielsen LB. 2004. Human placenta secretes apolipoprotein B-100-containing lipoproteins. *J Biol Chem* **279**: 55271-55276.

Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. 2019. Structural variant calling: the long and the short of it. *Genome Biol* **20**: 1-14.

Mannaerts GP, Van Veldhoven PP. 1996. Functions and organization of peroxisomal β-oxidation. *Ann N Y Acad Sci* **804**: 99-115.

Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. 2018. MUMmer4: a fast and versatile genome alignment system. *PLoS Comp Biol* **14**: e1005944.

Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**: 764-770.

Marsh-Matthews E, Deaton R. 2006. RESOURCES AND OFFSPRING PROVISIONING: A TEST OF THE TREXLER-DeANGELIS MODEL FOR MATROTROPHY EVOLUTION. *Ecology* **87**: 3014-3020.

Marsh-Matthews E, Skierkowski P, DeMarais A. 2001. Direct evidence for mother-to-embryo transfer of nutrients in the livebearing fish Gambusia geiseri. *Copeia* **2001**: 1-6.

Martin M, Patterson M, Garg S, Fischer S, Pisanti N, Klau GW, Schöenhuth A, Marschall T. 2016. WhatsHap: fast and accurate read-based phasing. *BioRxiv*: 085050.

Martinez N, Capobianco E, White V, Pustovrh M, Higa R, Jawerbaum A. 2008. Peroxisome proliferator-activated receptor a activation regulates lipid metabolism in the feto-placental unit from diabetic rats. *Reproduction* **136**: 95-104.

Mateos M, Kang D, Klopp C, Parrinello H, Garcia M, Schumer M, Jue N, Guiguen Y, Schartl M. 2019. Draft genome assembly and annotation of the Gila topminnow Poeciliopsis occidentalis. *Frontiers in Ecology and Evolution* **7**: 404.

McCarthy NS, Melton PE, Cadby G, Yazar S, Franchina M, Moses EK, Mackey DA, Hewitt AW. 2014. Meta-analysis of human methylation data for evidence of sex-specific autosomal patterns. *BMC Genomics* **15**: 1-11.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297-1303.

Meierjohann S, Schartl M. 2006. From Mendelian to molecular genetics: the *Xiphophorus* melanoma model. *Trends Genet* **22**: 654-661.

Miller W, Makova KD, Nekrutenko A, Hardison RC. 2004. Comparative genomics. *Annu Rev Genomics Hum Genet* **5**: 15-56.

Mishra A, Pokhrel P, Hoque MT. 2019. StackDPPred: a stacking based prediction of DNA-binding protein from sequence. *Bioinformatics* **35**: 433-441.

Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. 2013. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res* **41**: e121-e121.

Miyamori H, Takino T, Kobayashi Y, Tokai H, Itoh Y, Seiki M, Sato H. 2001. Claudin promotes activation of pro-matrix metalloproteinase-2 mediated by membrane-type matrix metalloproteinases. *J Biol Chem* **276**: 28204-28211.

Moccia F, Negri S, Shekha M, Faris P, Guerra G. 2019. Endothelial Ca2+ signaling, angiogenesis and vasculogenesis: just what it takes to make a blood vessel. *International journal of molecular sciences* **20**: 3962.

Moignard V, Woodhouse S, Haghverdi L, Lilly AJ, Tanaka Y, Wilkinson AC, Buettner F, Macaulay IC, Jawaid W, Diamanti E. 2015. Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat Biotechnol* **33**: 269-276.

Monk D. 2015. Genomic imprinting in the human placenta. *American journal of obstetrics and gynecology* **213**: S152-S162.

Nagai M, Oshima N, Fujii R. 1986. A comparative study of melanin-concentrating hormone (MCH) action on teleost melanophores. *The Biological Bulletin* **171**: 360-370.

Newburger DE, Bulyk ML. 2009. UniPROBE: an online database of protein binding microarray data on protein–DNA interactions. *Nucleic Acids Res* **37**: D77-D82.

Nie D-S, Liu Y, Juan H, Yang X. 2013. Overexpression of human SPATA17 protein induces germ cell apoptosis in transgenic male mice. *Mol Biol Rep* **40**: 1905-1910.

Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol* **3**: e170.

Nilsson D-E. 2013. Eye evolution and its functional basis. *Visual neuroscience* **30**: 5-20.

Nixon B, Johnston SD, Skerrett-Byrne DA, Anderson AL, Stanger SJ, Bromfield EG, Martin JH, Hansbro PM, Dun MD. 2018. Proteomic profiling of crocodile spermatozoa refutes the tenet that post-testicular maturation is restricted to mammals. *Molecular & Cellular Proteomics*: mcp. RA118. 000904.

O'Neill MJ, Lawton BR, Mateos M, Carone DM, Ferreri GC, Hrbek T, Meredith RW, Reznick DN, O'Neill RJ. 2007. Ancient and continuing Darwinian selection on insulin-like growth factor II in placental fishes. *Proceedings of the National Academy of Sciences* **104**: 12404-12409.

Oakley TH, Speiser DI. 2015. How complexity originates: the evolution of animal eyes. *Annual Review of Ecology, Evolution, and Systematics* **46**: 237-260.

Oda K, Shiratsuchi T, Nishimori H, Inazawa J, Yoshikawa H, Taketani Y, Nakamura Y, Tokino T. 1999. Identification of BAIAP2 (BAI-associated protein 2), a novel human homologue of hamster IRSp53, whose SH3 domain interacts with the cytoplasmic domain of BAI1. *Cytogenet Genome Res* **84**: 75-82.

Ohno S. 1970. The enormous diversity in genome sizes of fish as a reflection of nature's extensive experiments with gene duplication. *Trans Am Fish Soc* **99**: 120-130.

Olivera-Tlahuel C, Moreno-Mendoza NA, Villagrán-Santa Cruz M, Zúñiga-Vega JJ. 2018. Placental structures and their association with matrotrophy and superfetation in poeciliid fishes. *Acta Zoologica*.

Onodera T, Sakai T, Hsu JC-f, Matsumoto K, Chiorini JA, Yamada KM. 2010. Btbd7 regulates epithelial cell dynamics and branching morphogenesis. *Science* **329**: 562-565.

Ortiz-Zarragoitia M, Cajaraville MP. 2005. Effects of selected xenoestrogens on liver peroxisomes, vitellogenin levels and spermatogenic cell proliferation in male zebrafish. *Comp Biochem Physiol C Toxicol Pharmacol* **141**: 133-144.

Pang P-C, Chiu PC, Lee C-L, Chang L-Y, Panico M, Morris HR, Haslam SM, Khoo K-H, Clark GF, Yeung WS. 2011. Human sperm binding is mediated by the sialyl-Lewisx oligosaccharide on the zona pellucida. *Science* **333**: 1761-1764.

Parenti LR. 1981. A phylogenetic and biogeographic analysis of cyprinodontiform fishes (Teleostei, Atherinomorpha). *Bull Am Mus Nat Hist*.

Parenti LR, LoNostro FL, Grier HJ. 2010. Reproductive histology of Tomeurus gracilis Eigenmann, 1909 (Teleostei: Atherinomorpha: Poeciliidae) with comments on evolution of viviparity in atherinomorph fishes. *J Morphol* **271**: 1399-1406.

Park Y-J, Yoo S-A, Kim M, Kim W-U. 2020. The Role of Calcium–Calcineurin–NFAT Signaling Pathway in Health and Autoimmune Diseases. *Frontiers in immunology* **11**: 195.

Partha R, Chauhan BK, Ferreira Z, Robinson JD, Lathrop K, Nischal KK, Chikina M, Clark NL. 2017. Subterranean mammals show convergent regression in ocular genes and enhancers, along with adaptation to tunneling. *Elife* **6**: e25884.

Paten B, Earl D, Nguyen N, Diekhans M, Zerbino D, Haussler D. 2011. Cactus: Algorithms for genome multiple sequence alignment. *Genome Res* **21**: 1512-1528.

Pina C, Teles J, Fugazza C, May G, Wang D, Guo Y, Soneji S, Brown J, Edén P, Ohlsson M. 2015. Single-cell network analysis identifies DDIT3 as a nodal lineage regulator in hematopoiesis. *Cell reports* **11**: 1503-1510.

Piot A, Hackel J, Christin P-A, Besnard G. 2018. One-third of the plastid genes evolved under positive selection in PACMAD grasses. *Planta* **247**: 255-266.

Pollux B, Meredith R, Springer M, Garland T, Reznick D. 2014. The evolution of the placenta drives a shift in sexual selection in livebearing fish. *Nature* **513**: 233.

Pollux B, Pires M, Banet A, Reznick D. 2009. Evolution of placentas in the fish family Poeciliidae: an empirical study of macroevolution. *Annu Rev Ecol Evol Syst* **40**: 271-289.

Pollux BJ, Reznick DN. 2011. Matrotrophy limits a female's ability to adaptively adjust offspring size and fecundity in fluctuating environments. *Funct Ecol* **25**: 747-756.

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

Pryszcz LP, Gabaldón T. 2016. Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Res* **44**: e113-e113.

Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**: i333-i339.

Razin A, Cedar H. 1991. DNA methylation and gene expression. *Microbiol Mol Biol Rev* **55**: 451-458.

Reddy JK, Chu R. 1996. Peroxisome Proliferator–induced Pleiotropic Responses: Pursuit of a Phenomenon a. *Ann N Y Acad Sci* **804**: 176-201.

Reik W, Walter J. 2001. Genomic imprinting: parental influence on the genome. *Nature Reviews Genetics* **2**: 21.

Reis RE, Kullander SO, Ferraris CJ. 2003. *Check list of the freshwater fishes of South and Central America*. Edipucrs.

Revell LJ. 2012. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in ecology and evolution* **3**: 217-223.

Reznick DN, Furness AI, Meredith RW, Springer MS. 2017. The origin and biogeographic diversification of fishes in the family Poeciliidae. *PloS one* **12**: e0172546.

Reznick DN, Mateos M, Springer MS. 2002. Independent origins and rapid evolution of the placenta in the fish genus *Poeciliopsis*. *Science* **298**: 1018-1020.

Reznick DN, Rodd FH, Cardenas M. 1996. Life-history evolution in guppies (*Poecilia reticulata*: Poeciliidae). IV. Parallelism in life-history phenotypes. *The American Naturalist* **147**: 319-338.

Rhoads A, Au KF. 2015. PacBio sequencing and its applications. *Genomics, proteomics & bioinformatics* **13**: 278-289.

Riesch R, Martin RA, Langerhans RB. 2012. Predation's role in life-history evolution of a livebearing fish and a test of the Trexler-DeAngelis model of maternal provisioning. *The American Naturalist* **181**: 78-93.

Rocha MJ, Arukwe A, Kapoor B. 2008. *Fish reproduction*. CRC Press.

Rose AJ, Kiens B, Richter EA. 2006. Ca2+–calmodulin-dependent protein kinase expression and signalling in skeletal muscle during exercise. *The Journal of physiology* **574**: 889-903.

Ross BD, Rosin L, Thomae AW, Hiatt MA, Vermaak D, de la Cruz AFA, Imhof A, Mellone BG, Malik HS. 2013. Stepwise evolution of essential centromere function in a Drosophila neogene. *Science* **340**: 1211-1214.

Rossant J, Cross JC. 2001. Placental development: lessons from mouse mutants. *Nature Reviews Genetics* **2**: 538.

Roth O, Solbakken MH, Tørresen OK, Bayer T, Matschiner M, Baalsrud HT, Hoff SNK, Brieuc MSO, Haase D, Hanel R. 2020. Evolution of male pregnancy associated with remodeling of canonical vertebrate immunity in seahorses and pipefishes. *Proc Natl Acad Sci* **117**: 9431-9439.

Ruan J, Li H. 2019. Fast and accurate long-read assembly with wtdbg2. *BioRxiv*: 530972.

Ruan J, Li H. 2020. Fast and accurate long-read assembly with wtdbg2. *Nat Methods* **17**: 155-158.

Sackton TB, Grayson P, Cloutier A, Hu Z, Liu JS, Wheeler NE, Gardner PP, Clarke JA, Baker AJ, Clamp M. 2019. Convergent regulatory evolution and loss of flight in paleognathous birds. *Science* **364**: 74-78.

Saegusa J, Akakura N, Wu C-Y, Hoogland C, Ma Z, Lam KS, Liu F-T, Takada YK, Takada Y. 2008. Pro-inflammatory secretory phospholipase A2 type IIA binds to integrins αvβ3 and α4β1 and induces proliferation of monocytic cells in an integrin-dependent manner. *J Biol Chem* **283**: 26107-26115.

Sapiro R, Kostetskii I, Olds-Clarke P, Gerton GL, Radice GL, Strauss III JF. 2002. Male infertility, impaired sperm motility, and hydrocephalus in mice deficient in sperm-associated antigen 6. *Mol Cell Biol* **22**: 6298-6305.

Sato T, Yamanishi Y, Kanehisa M, Toh H. 2005. The inference of protein–protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. *Bioinformatics* **21**: 3482-3489.

Schartl M, Walter RB, Shen Y, Garcia T, Catchen J, Amores A, Braasch I, Chalopin D, Volff J-N, Lesch K-P. 2013. The genome of the platyfish, *Xiphophorus maculatus*, provides insights into evolutionary adaptation and several complex traits. *Nat Genet* **45**: 567.

Schrader M, Travis J. 2012. Embryonic IGF2 expression is not associated with offspring size among populations of a placental fish. *PloS one* **7**: e45463.

Sessegolo C, Cruaud C, Da Silva C, Cologne A, Dubarry M, Derrien T, Lacroix V, Aury J-M. 2019. Transcriptome profiling of mouse samples using nanopore sequencing of cDNA and RNA molecules. *Scientific reports* **9**: 1-12.

Shen Y, Chalopin D, Garcia T, Boswell M, Boswell W, Shiryev SA, Agarwala R, Volff J-N, Postlethwait JH, Schartl M. 2016a. *X. couchianus* and *X. hellerii* genome models provide genomic variation insight among *Xiphophorus* species. *BMC Genomics* **17**: 37.

Shen Y, Chalopin D, Garcia T, Boswell M, Boswell W, Shiryev SA, Agarwala R, Volff J-N, Postlethwait JH, Schartl M. 2016b. X. couchianus and X. hellerii genome models provide genomic variation insight among Xiphophorus species. *BMC Genomics* **17**: 1-13.

Shen Y, Ge W-P, Li Y, Hirano A, Lee H-Y, Rohlmann A, Missler M, Tsien RW, Jan LY, Fu Y-H. 2015. Protein mutated in paroxysmal dyskinesia interacts with the active zone protein RIM and suppresses synaptic vesicle exocytosis. *Proceedings of the National Academy of Sciences* **112**: 2935-2941.

Shoemaker R, Deng J, Wang W, Zhang K. 2010. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. *Genome Res* **20**: 883-889.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**: 3210-3212.

Simpson J. 2018. Nanopolish: Signal-level algorithms for MinION data. *Github Available at: https://github com/jts/nanopolish [Accessed January 10, 2019]*.

Simpson JT, Workman RE, Zuzarte P, David M, Dursi L, Timp W. 2017. Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods* **14**: 407-410.

Sirén J, Välimäki N, Mäkinen V. 2014. HISAT2-fast and sensitive alignment against general human population. *IEEE/ACM Trans Comput Biol Bioinforma* **11**: 375-388.

Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH. 2009a. JBrowse: a next-generation genome browser. *Genome Res*: gr. 094607.094109.

Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH. 2009b. JBrowse: a next-generation genome browser. *Genome Res* **19**: 1630-1638.

Slater GSC, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**: 31.

Smit A, Hubley R. 2008. RepeatModeler Open-1.0. *Available fom http://www repeatmasker org*.

Smit A, Hubley R, Green P. 2013. RepeatMasker Open-4.0. *Available from: http://wwwrepeatmaskerorg*.

Sonderegger S, Pollheimer J, Knöfler M. 2010. Wnt signalling in implantation, decidualisation and placental differentiation–review. *Placenta* **31**: 839-847.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312-1313.

Stemmler MP, Bedzhov I. 2010. A Cdh1HA knock-in allele rescues the Cdh1−/− phenotype but shows essential Cdh1 function during placentation. *Dev Dyn* **239**: 2330-2344.

Studer RA, Penel S, Duret L, Robinson-Rechavi M. 2008. Pervasive positive selection on duplicated and nonduplicated vertebrate protein coding genes. *Genome Res* **18**: 1393-1402.

Subramanian S, Mishra RK, Singh L. 2003. Genome-wide analysis of Bkm sequences (GATA repeats): predominant association with sex chromosomes and potential role in higher order chromatin organization and function. *Bioinformatics* **19**: 681-685.

Summers K, Crespi B. 2005. Cadherins in maternal–foetal interactions: red queen with a green beard? *Proceedings of the Royal Society B: Biological Sciences* **272**: 643-649.

Szabova L, Son M-Y, Shi J, Sramko M, Yamada SS, Swaim WD, Zerfas P, Kahan S, Holmbeck K. 2010. Membrane-type MMPs are indispensable for placental labyrinth formation and development. *Blood* **116**: 5752-5761.

Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva NT, Roth A, Bork P. 2016. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res*: gkw937.

Tak Leung RW, Jiang X, Chu KH, Qin J. 2019. ENPD-a database of eukaryotic nucleic acid binding proteins: Linking gene regulations to proteins. *Nucleic Acids Res* **47**: D322-D329.

Tarver JE, Dos Reis M, Mirarab S, Moran RJ, Parker S, O'Reilly JE, King BL, O'Connell MJ, Asher RJ, Warnow T. 2016. The interrelationships of placental mammals and the limits of phylogenetic inference. *Genome biology and evolution* **8**: 330-344.

Teng L, Fan X, Xu D, Zhang X, Mock T, Ye N. 2017. Identification of genes under positive selection reveals differences in evolutionary adaptation between brown-algal species. *Frontiers in plant science* **8**: 1429.

Thibodeau A, Uyar A, Khetan S, Stitzel ML, Ucar D. 2018. A neural network based model effectively predicts enhancers from clinical ATAC-seq samples. *Scientific reports* **8**: 1-15.

Thiel T, Michalek W, Varshney R, Graner A. 2003. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). *Theor Appl Genet* **106**: 411-422.

Torgerson DG, Kulathinal RJ, Singh RS. 2002. Mammalian sperm proteins are rapidly evolving: evidence of positive selection in functionally diverse genes. *Mol Biol Evol* **19**: 1973-1980.

Trexler JC. 1985. Variation in the degree of viviparity in the sailfin molly, Poecilia latipinna. *Copeia*: 999-1004.

Trexler JC. 1997. Resource availability and plasticity in offspring provisioning: embryo nourishment in sailfin mollies. *Ecology* **78**: 1370-1381.

Trexler JC, Travis J, Trexler M. 1990. Phenotypic plasticity in the sailfin molly, *Poecilia latipinna* (Pisces: Poeciliidae). II. Laboratory experiment. *Evolution* **44**: 157-167.

Trivers RL. 1974. Parent-offspring conflict. *Integr Comp Biol* **14**: 249-264.

Tucci V, Isles AR, Kelsey G, Ferguson-Smith AC, Bartolomei MS, Benvenisty N, Bourc'his D, Charalambous M, Dulac C, Feil R. 2019. Genomic imprinting and physiological processes in mammals. *Cell* **176**: 952-965.

Turner CL. 1940. Pseudoamnion, pseudochorion, and follicular pseudoplacenta in poeciliid fishes. *J Morphol* **67**: 59-89.

Twiss F, de Rooij J. 2013. Cadherin mechanotransduction in tissue remodeling. *Cell Mol Life Sci* **70**: 4101-4116.

Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A. 2015. Tissue-based map of the human proteome. *Science* **347**: 1260419.

Van Der Laan R, Eschmeyer WN, Fricke R. 2014. Family-group names of recent fishes. *Zootaxa* **3882**: 1-230.

van Kruistum H, Guernsey MW, Baker JC, Kloet SL, Groenen MA, Pollux BJ, Megens H-J. 2020. The genomes of the livebearing fish species Poeciliopsis retropinna and Poeciliopsis turrubarensis reflect their different reproductive strategies. *Mol Biol Evol* **37**: 1376-1386.

van Kruistum H, Nijland R, Reznick DN, Groenen MA, Megens H-J, Pollux BJ. 2021. Parallel genomic changes drive repeated evolution of placentas in live-bearing fish. *Mol Biol Evol*.

Van Kruistum H, Van Den Heuvel J, Travis J, Kraaijeveld K, Zwaan BJ, Groenen MA, Megens H-J, Pollux BJ. 2019. The genome of the live-bearing fish Heterandria formosa implicates a role of conserved vertebrate genes in the evolution of placental fish. *BMC Evol Biol* **19**: 156.

Vardimon L, Kressmann A, Cedar H, Maechler M, Doerfler W. 1982. Expression of a cloned adenovirus gene is inhibited by in vitro methylation. *Proceedings of the National Academy of Sciences* **79**: 1073-1077.

Vega-López A, Ortiz-Ordóñez E, Uría-Galicia E, Mendoza-Santana EL, Hernández-Cornejo R, Atondo-Mexia R, García-Gasca A, García-Latorre E, Domínguez-López ML. 2007. The role of vitellogenin during gestation of *Girardinichthys viviparus* and *Ameca splendens*; two goodeid fish with matrotrophic viviparity. *Comp Biochem Physiol A Mol Integr Physiol* **147**: 731-742.

Wagner A. 1994. Evolution of gene networks by gene duplications: a mathematical model and its implications on genome organization. *Proceedings of the National Academy of Sciences* **91**: 4387-4391.

Wagner A. 2008. Gene duplications, robustness and evolutionary innovations. *Bioessays* **30**: 367-373.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one* **9**: e112963.

Wang RN, Green J, Wang Z, Deng Y, Qiao M, Peabody M, Zhang Q, Ye J, Yan Z, Denduluri S. 2014. Bone Morphogenetic Protein (BMP) signaling in development and human diseases. *Genes & diseases* **1**: 87-105.

Wang W-J, Tay HG, Soni R, Perumal GS, Goll MG, Macaluso FP, Asara JM, Amack JD, Tsou M-FB. 2013. CEP162 is an axoneme-recognition protein promoting ciliary transition zone assembly at the cilia base. *Nat Cell Biol* **15**: 591.

Wang YR, Li L, Li JJ, Huang H. 2021. Network modeling in biology: statistical methods for gene and brain networks. *Statistical Science* **36**: 89-108.

Warren WC, García-Pérez R, Xu S, Lampert KP, Chalopin D, Stöck M, Loewe L, Lu Y, Kuderna L, Minx P. 2018. Clonal polymorphism and high heterozygosity in the celibate genome of the Amazon molly. *Nature ecology & evolution* **2**: 669.

Watson JD, Crick FH. 1953. The structure of DNA. In *Cold Spring Harbor Symp Quant Biol*, Vol 18, pp. 123-131. Cold Spring Harbor Laboratory Press.

Wei Z-B, Yuan Y-F, Jaouen F, Ma M-S, Hao C-J, Zhang Z, Chen Q, Yuan Z, Yu L, Beurrier C. 2016. SLC35D3 increases autophagic activity in midbrain dopaminergic neurons by enhancing BECN1-ATG14-PIK3C3 complex formation. *Autophagy* **12**: 1168-1179.

Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. 2017. Direct determination of diploid genome sequences. *Genome Res* **27**: 757-767.

Wetterstrand KA. 2021. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP).

Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol* **18**: 691-699.

Wilkins JF, Haig D. 2003. What good is genomic imprinting: the function of parent-specific gene expression. *Nature Reviews Genetics* **4**: 359-368.

Wolverton T, Lalande M. 2001. Identification and characterization of three members of a novel subclass of protocadherins. *Genomics* **76**: 66-72.

Workman RE, Tang AD, Tang PS, Jain M, Tyson JR, Razaghi R, Zuzarte PC, Gilpatrick T, Payne A, Quick J. 2019. Nanopore native RNA sequencing of a human poly (A) transcriptome. *Nat Methods* **16**: 1297-1305.

Wourms JP, Grove BD, Lombardi J. 1988. 1 The Maternal-Embryonic Relationship in Viviparous Fishes. In *Fish physiology*, Vol 11, pp. 1-134. Elsevier.

Yang Q, Wang H-X, Zhao Y-G, Lin H-Y, Zhang H, Wang H-M, Sang Q-XA, Zhu C. 2006. Expression of tissue inhibitor of metalloproteinase-4 (TIMP-4) in endometrium and placenta of rhesus monkey (Macaca mulatta) during early pregnancy. *Life Sci* **78**: 2804-2811.

Yang TT, Raymond Y, Agadir A, Gao G-J, Campos-Gonzalez R, Tournier C, Chow C-W. 2008. Integration of protein kinases mTOR and extracellular signal-regulated kinase 5 in regulating nucleocytoplasmic localization of NFATc4. *Mol Cell Biol* **28**: 3489-3501.

Yang X, Han H, De Carvalho DD, Lay FD, Jones PA, Liang G. 2014. Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer Cell* **26**: 577-590.

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

Ye K, Guo L, Yang X, Lamijer E-W, Raine K, Ning Z. 2018. Split-read indel and structural variant calling using PINDEL. In *Copy Number Variants*, pp. 95-105. Springer.

Yu D, Kim M, Xiao G, Hwang TH. 2013. Review of biological network data and its applications. *Genomics & informatics* **11**: 200.

Zandonà E, Kajin M, Buckup PA, Amaral JR, Souto-Santos IC, Reznick DN. 2021. Mode of maternal provisioning in the fish genus Phalloceros: a variation on the theme of matrotrophy. *Biol J Linn Soc*.

Zartman JJ, Shvartsman SY. 2010. Unit operations of tissue development: epithelial folding. *Annual review of chemical and biomolecular engineering* **1**: 231-246.

Zeh DW, Zeh JA. 2000. Reproductive mode and speciation: the viviparity-driven conflict hypothesis. *Bioessays* **22**: 938-946.

Zeh JA, Zeh DW. 2001. Reproductive mode and the genetic benefits of polyandry. *Anim Behav* **61**: 1051-1063.

Zhang C, Sayyari E, Mirarab S. 2017. ASTRAL-III: increased scalability and impacts of contracting low support branches. In *RECOMB international workshop on comparative genomics*, pp. 53-75. Springer.

Zhang Y, Shin H, Song JS, Lei Y, Liu XS. 2008. Identifying positioned nucleosomes with epigenetic marks in human from ChIP-Seq. *BMC Genomics* **9**: 537.

Zhang Z, Zariwala MA, Mahadevan MM, Caballero-Campo P, Shen X, Escudier E, Duriez B, Bridoux A-M, Leigh M, Gerton GL. 2007. A heterozygous mutation disrupting the SPAG16 gene results in biochemical instability of central apparatus components of the human sperm axoneme. *Biol Reprod* **77**: 864-871.

Zhao G-Q, Hogan BL. 1996. Evidence that mouse Bmp8a (Op2) and Bmp8b are duplicated genes that play a role in spermatogenesis and placental development. *Mech Dev* **57**: 159-168.

Zheng Q, Zheng X, Zhang L, Luo H, Qian L, Fu X, Liu Y, Gao Y, Niu M, Meng J. 2017. The Neuron-Specific Protein TMEM59L Mediates Oxidative Stress-Induced Cell Death. *Mol Neurobiol* **54**: 4189-4200.

Summary

All life that we know develops according to instructions within its DNA. As organisms evolve, changes in an their body plan should therefore be reflected in associated changes in their genome. However, finding the mutations that cause complex phenotypic change is still a great challenge. In this thesis, I aim to find the genomic changes associated with the evolution of the placenta in the livebearing fish family Poeciliidae. For this, I sequenced, assembled, and then compared the genomes of 26 poeciliid species, aiming to find consistent genomic differences between placental and non-placental species.

In **chapter 2**, I re-assembled the genome of the placental livebearer *Heterandria formosa*, using existing sequencing data, after which I compared it to the publicly available genomes of three non-placental relatives. I show that a number of genes in the genome of *H. formosa* that are related to placenta formation show signs of positive selection, while evolving under regular constraint in the non-placental species. Additionally, I identify a small number of gene duplications that are unique to *H. formosa*.

Chapter 3 describes the genome assemblies of two poeciliid species for which the genome sequence was previously unknown: the placental *Poeciliopsis retropinna* and the non-placental *Poeciliopsis turrubarensis*. With the third-generation sequencing techniques that were used, assemblies of excellent quality could be generated. I used these assemblies to reliably identify structural variations between the two species, with the tandem duplication of the *vtg1* gene in *P. retropinna* being a particularly interesting result.

In **chapter 4** I compare the genomes of 26 poeciliid species, both publicly available as well as assembled in this chapter or in previous chapters. I show that placental species within this group display accelerated evolution of genes involved with transporter- and vesicle-related functions, providing first evidence for genomic convergence associated with placenta evolution. Additionally, I observed differences in the presence of regulatory elements around developmental genes that were not observed in non-placental species, indicating that regulatory change is also a part of the evolution of the poeciliid placenta.

In **Chapter 5** I investigate the occurrence of gene duplications and deletions across the genomes of twelve poeciliid species, three of which are placental. For this, I develop a new pipeline that can simultaneously identify structural variants in the genomes of multiple related species. According to this analysis, placenta evolution in the Poeciliidae is not associated with gene duplications, but instead gene deletions were found in the same molecular pathways for the three placental species: the calcineurin-NFAT signaling cascade and the BMP signaling pathway. In non-placental species, these deletions did not occur.

Finally, **Chapter 6** describes a new approach to detect allele-specific methylation based on Oxford Nanopore sequencing. I apply this approach to four individuals of the non-placental *Poeciliopsis gracilis*: two parents, and a male and a female offspring. I show that allele-specific methylation is widespread in the genome of *P. gracilis*. Additionally, the inheritance of methylated alleles is not always random, but instead depends on parent-of-origin. The genes that are in the vicinity of regions that are affected by parent-specific methylation are located predominantly in the brain. These results lead to the hypothesis that genomic imprinting due to intralocus sexual conflict is the cause of the observed allele-specific methylation.

Curriculum Vitae

About the author

Hendrik van Kruistum, who we all know as Henri, was born in the Netherlands in Amersfoort on Friday October 28th, 1994. He was raised with his two younger sisters in the small town Kootwijkerbroek, northeast of Barneveld.

During his childhood, Henri was already a little bit of a scientist. He was very inquisitive and an avid reader, learning anything he could about our solar system and its planets. After learning as much as he could about one topic, he would move on to the next. In high school, biology became one of Henri's favorite topics. After finishing high school, Henri started his bachelor and subsequently master Biotechnology at Wageningen University. Although the focus of his minor thesis, major thesis and internship was about bacteria, Henri thought it was a good idea to expand his horizons and work on a topic he was a little less familiar with. A PhD was the next step in his career, the topic was going to be the most interesting vacancy he could find.

Being the owner of a fish tank during most of his childhood and adult life (with its ups and downs; at some point in time his tank got the nickname algae paradise), it was an obvious choice for Henri to apply for the PhD on the topic of live-bearing fish. Bioinformatics was something he had dabbled in a bit, but it was a subject he was eager to learn. After having talked enthusiastically about his fish tank during his job interviews, his supervisors must have thought Henri was an obvious match for this PhD as well.

Fast forward to 4 years later, and Henri is back to his old tricks. Although staying on-topic by applying to a job as bioinformatician, Henri will be exploring the field of plants next. We are very excited to see where this path will take him.

Peer-reviewed Publications

van Kruistum, H., Nijland, R., Reznick, D. N., Groenen, M., Pollux, B. J. A., & Megens, H. J. W. C. (2020). Parallel genomic changes drive repeated evolution of placentas in livebearing fish. Molecular biology and evolution, 38 (6), 2627-2638.

van Kruistum, H., Guernsey, M. W., Baker, J. C., Kloet, S. L., Groenen, M. A., Pollux, B. J., & Megens, H. J. (2020). The genomes of the livebearing fish species *Poeciliopsis retropinna* and *Poeciliopsis turrubarensis* reflect their different reproductive strategies. Molecular biology and evolution, 37(5), 1376-1386.

Guernsey, M. W., **van Kruistum, H**., Reznick, D. N., Pollux, B. J., & Baker, J. C. (2020). Molecular signatures of placentation and secretion uncovered in *Poeciliopsis* maternal follicles. Molecular Biology and Evolution 37(9), 2679–2690.

Van Kruistum, H., Van Den Heuvel, J., Travis, J., Kraaijeveld, K., Zwaan, B. J., Groenen, M. A., Megens, H. J. & Pollux, B. J. (2019). The genome of the live-bearing fish *Heterandria formosa* implicates a role of conserved vertebrate genes in the evolution of placental fish. BMC evolutionary biology, 19(1), 156.

van Kruistum, H., Bodelier, P. L., Ho, A., Meima-Franke, M., & Veraart, A. J. (2018). Resistance and recovery of methane-oxidizing communities depends on stress regime and history; a microcosm study. Frontiers in microbiology, 9, 1714.

Strepis, N., Sánchez-Andrea, I., van Gelder, A. H., **van Kruistum, H**., Shapiro, N., Kyrpides, N., ... & Sousa, D. Z. (2016). Description of *Trichococcus ilyis sp. nov.* by combined physiological and in silico genome hybridization analyses. International journal of systematic and evolutionary microbiology, 66(10), 3957-3963

Under review

van kruistum, H., Nijland, R., Ernst, T.R., Madsen, O., Groenen, M.A.M., Pollux, B.J.A.P., Megens, H-J. Allele-specific methylation in the livebearing fish *Poeciliopsis gracilis*.

van kruistum, H., Groenen, M.A.M, Megens, H-J., Pollux, B.J.A.P. Not gene duplications, but gene deletions are associated with placenta evolution in livebearing fish.

Training and Education

The Basic Package (1.8 credits)	
WIAS Introduction Day (0.3 credits)	2018
Course on philosophy of science and/or ethics (1.5 credits)	2018

Disciplinary Competences (13.5 credits)	
Writing a literature survey (6 credits)	2018
Physalia course: Genomic signatures of selection (3 credits)	2019
PhD discussion group (ABG weekly genomics meeting) (1.5 credits)	2018-2021
BioSB course: Algorithms for biological networks (3 credits)	2021

Professional Competences (6.5 credits)	
Research data management (0.5 credits)	2018
Brain training (0.3 credits)	2018
Joining the WAPS council: 1 year accie, 1 year WPC representative (3 credits)	2018-2020
WGS course: Scientific artwork, data visualisation and infographics with Adobe Illustrator (0.6 credits)	2020
WGS course: Career orientation (1.5 credits)	2021
WIAS course: the final touch (0.6 credits)	2021

Societal Relevance (1.5 credits)	
Societal Impact of your Research (1.5 credits)	2020

Presentation Skills (4.0 credits)	
Netherlands Annual Ecology Meeting (oral)	2019
European Society for Evolutionary Biology Congress (poster)	2019
WIAS annual conference (oral)	2020
Nanopore Community Meeting (oral)	2021

Teaching competences (6.0 credits)	
Supervising Genomics (ABG-30306) four times total (4 credits)	2019-2020
Supervising MSc Major thesis Nathalie de Vries (2 credits)	2019-2020

Total credits	**33**

Acknowledgements

A PhD thesis is not completed alone. Several people have helped me along the way either directly, or by helping me become the person that I am today. I would like to write a few sentences to thank them.

First, I would like to thank my daily supervisors, Bart and Hendrik-Jan. When I started this thesis, I knew almost nothing about both fish and bioinformatics, but both of you are extremely knowledgeable in your field and were always willing to answer my questions, allowing me to learn quickly. Besides that, both of you gave me the freedom to do my own thing and work independently, which I am grateful for. I enjoyed our coffee breaks, where we discussed many things but only rarely talked about this project. I think our collaboration has been very successful. Although not officially a supervisor, I feel that I also have to thank Reindert here. Reindert, thank you for your help with everything related to Nanopore sequencing.

Then, I would like to take a moment to thank my paranymphs, Michael and Renzo. Renzo, together we achieved the largest triumph attainable by winning the ABGC Foosball Champion Challenge Cup, an honour only awarded to the absolute elite in the competitive field of foosball connoisseurs. Thank you for your memes and friendship. Michael, you always had great ideas for activities and I enjoyed all of them, whether it were foosball, beers, RISK, or other games. The chess meetings were one of the few online meetings that I actually attended. Thank you for a great time.

I would also like to thank my colleagues, both from ABG and EZO. My friends from ABG, we had a good time drinking beers in De Vlaamsche Reus, during the Palm festival, or anywhere else. Sadly, Covid threw a wrench in many of these activities. Still, we did some cool stuff like playing RISK and the weekly chess meetings. I had a great time with you all. EZO-ites, I did not immediately get to know you very well as I worked mostly from Radix, but this changed after we started playing d&d. I have been enjoying these game nights a lot, and hope to continue playing for a long time!

Als laatste wil ik mijn vrouw, familie en vrienden bedanken. Laura, ik leerde je kennen vlak voor de start van mijn PhD, en nu zijn we getrouwd. Bij jou voel ik me thuis. Bedankt voor je steun. Paps en mams, bedankt voor jullie gezelligheid, bezorgdheid en rundvleessoep. Evithe en Ellen, bedankt voor de gezelligheid op de game zondagen en de potjes Catan. Vrienden van de Ulcus, nog bij ons of niet meer, bedankt voor de gouden tijden. Ik hoop op meer weekendjes weg in de toekomst.

Colophon

This research was funded by a WIAS research grant, and by the ABG and EZO research groups (Wageningen UR).

Printed by DigiForce | Proefschriftmaken.nl

ISBN: 978-94-6447-037-6
DOI: https://doi.org/10.18174/55808