UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS

Gene expression regulation in allopolyploid fish

"Documento Definitivo"

DOUTORAMENTO EM BIOLOGIA
ESPECIALIDADE EM BIOLOGIA MOLECULAR

Isa Maria Nunes de Matos

Tese orientada por:
Professora Doutora Maria Manuela Coelho
Professor Doutor Manfred Schartl

Documento especialmente elaborado para a obtenção do grau de doutor

2018

UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

Gene expression regulation in allopolyploid fish

DOUTORAMENTO EM BIOLOGIA

ESPECIALIDADE EM BIOLOGIA MOLECULAR

Isa Maria Nunes de Matos

Tese orientada por:

Professora Doutora Maria Manuela Coelho
Professor Doutor Manfred Schartl

Júri:

Presidente:

- Doutor Rui Manuel Dos Santos Malhó, Professor Catedrático, Faculdade de Ciências da Universidade de Lisboa

Vogais:

- Doutor Manfred Schartl, *Full Professor*, *Biology Faculty* da *University of Wuerzburg*, Alemanha (Orientador)
- Doutor Vasco Temudo e Melo Cabral Barreto, Investigador Principal, Centro de Estudos de Doenças Crónicas da Universidade Nova de Lisboa
- Doutora Carla Patrícia Cândido de Sousa Santos, Investigadora, ISPA – Instituto Universitário de Doenças Psicológicas, Sociais e da Vida
- Doutor Rui Miguel Duque de Brito, Professor Coordenador com Agregação, Escola Superior de Tecnologias de Saúde de Lisboa do Instituto Politécnico de Lisboa
- Doutora Maria João Ivens Collares Pereira, Professora Catedrática Aposentada, Faculdade de Ciências da Universidade de Lisboa
- Doutor Vitor Martins Conde e Sousa, Professor Auxiliar Convidado, Faculdade de Ciências da Universidade de Lisboa

Documento especialmente elaborado para a obtenção do grau de doutor

Bolsa Individual de Doutoramento SFRH/BD/61217/2009
Fundação para a Ciência e a Tecnologia

2018

Nota prévia

Para a elaboração da presente dissertação, e nos termos do nº 2 do artigo 25º do Regulamento de Estudos de Pós-Graduação da Universidade de Lisboa, publicado no Diário da República, 2.ª série N.º 155 — 11 de Agosto de 2017, foram usados integralmente artigos científicos publicados (4) ou submetidos para publicação (1) em revistas internacionais indexadas. Tendo os trabalhos referidos sido efetuados em colaboração, a autora da dissertação esclarece que participou integralmente na conceção e execução do trabalho experimental, na análise e interpretação dos resultados e na redação de todos os manuscritos, mesmo do único em que não é primeira autora.

Lisboa, Outubro de 2018

Isa Maria Nunes de Matos

ACKNOWLEDGEMENTS

Foram muitas as pessoas que ao longo destes anos fizeram parte desta viagem. A todas agradeço o papel que tiveram. Quero, no entanto, agradecer especialmente a algumas delas.

À Professora Doutora Maria Manuela Coelho pelo apoio, confiança, motivação, paciência e carinho. Por me ter recebido no seu laboratório e ter estado sempre ao meu lado a festejar as pequenas vitorias e a amparar as quedas. Agradeço também por me ter confiado o "seu alburnoides", o seu projeto e as suas ideias, mas deixando sempre espaço para as minhas iniciativas e planos menos convencionais.

To Professor Doctor Manfred Schartl, for the scientific guidance and support, always. For opening the doors of his lab, make available his resources and knowledge to make this project happen. Thanks for the patience and understanding of my many flaws. It was a personal and professional privilege to work and learn from you.

À Professora Maria João Collares-Pereira, pela disponibilidade em partilhar todo o seu enorme conhecimento sobre este fascinante modelo que é o *S. alburnoides*, sempre com simpatia e boa disposição.

Aos "Meus Migueis", quero agradecer a loucura do dia-dia, a emoção e a festa que foram estes anos partilhados. Sem vocês… não sei….

À minha Su, que melhor que ninguém sabe como isto é. Contigo por perto foi mais fácil. Agora é a tua vez!

À Tia Inês, que apesar de ser de "letras" (ninguém é perfeito!) sempre se interessou pelo meu trabalho. Obrigado por teres ouvido as minhas "histórias dos peixes".

À minha família (em especial à minha Mãe, ao meu Pai, à Ana ao Pedro e à Avó). Vocês são as bases de tudo.

Ao meu Ricardo, o companheiro de todas as horas! Hoje e sempre "Apaga as luzes"!

Resumo

Os indivíduos resultantes de hibridação são muitas vezes inviáveis ou inférteis, pelo que, durante décadas, foram vistos como ocorrências pouco frequentes e meros erros irrelevantes para o processo evolutivo. No entanto, esta perspectiva redutora tem vindo progressivamente a ser abandonada e actualmente, a hibridação e a poliploidia, fenómenos que aparecem muitas vezes associados, são aceites como processos evolutivamente relevantes.

Em relação ás plantas, a ocorrência de (alo)poliploides dentro deste grupo é bastante frequente, e estas sempre foram reconhecidas como sendo bastante tolerantes a alterações de ploidia e à hibridação. Assim, os efeitos destes processos na expressão génica e na regulação das cópias de cada genoma têm sido amplamente estudados neste grupo. No entanto, a existência e importância da (alo)poliploidia em animais foi, por muitos anos, negligenciada e consequentemente, pouco se sabe sobre os seus efeitos na expressão génica e nos mecanismos da sua regulação em vertebrados (allo)poliplóides.

Posto isto, esta tese apresenta-se, como um avanço na informação disponível sobre o tema da regulação da expressão génica em animais, contribuindo para desvendar as causas e os mecanismos que levam a que alguns genomas poliploides híbridos superem eficazmente o "choque genómico" resultante do aumento do número de cromossomas e da combinação de genomas divergentes. Os estudos aqui englobados são também a continuação e o amadurecimento de descobertas e resultados anteriores, uma vez que a sua génese baseou-se numa interessante teoria de "diploidização funcional" da expressão génica em alotriploides do complexo *Squalius alburnoides* mediante silenciamento alélico. Este complexo de peixes alopoliploides, endémico da Península Ibérica, foi assim o primeiro modelo a ser estudado neste âmbito, pois apresenta características que o tornam particularmente interessante. O complexo resulta da hibridação interespecífica e não recíproca entre fêmeas de *Squalius pyrenaicus* (genoma P) e de uma espécie que se pensar estar actualmente extinta, próxima de *Anaecypris hispanica* (genoma

A). Inclui formas diploides e poliploides mantidas por diferentes modos de reprodução potenciando um sistema de trocas genéticas em que as diferentes formas diploides e poliploides se cruzam e contribuem activamente para a manutenção da diversidade e do potencial evolutivo da espécie. Apresenta ainda como caraterística um acentuado desvio na distribuição dos sexos, com uma predominância de fêmeas triplóides e ainda a ocorrência de uma linhagem aparentemente constituída exclusivamente por machos.

O objectivo geral deste trabalho foi ajudar a esclarecer como se processa a expressão génica em alotriploides de *S. alburnoides*, bem como de mecanismos ligados à sua regulação. Em concreto, pretendeu-se esclarecer se os mecanismos de compensação da dosagem e o silenciamento alélico, anteriormente descritos recorrendo à análise de um número muito limitado de genes, são fenómenos com extensão genómica ou se estão limitados, específica ou aleatoriamente, a subgrupos de genes. No que respeita ao silenciamento alélico, pretendeu-se ainda esclarecer se existe enviesamento da expressão, favorecendo um ou outro complemento genómico, e, se sim, perceber a sua magnitude e a sua extensão no genoma. Sendo o silenciamento alélico o limite máximo do enviesamento da expressão alélica, e também o mecanismo de compensação de dosagem anteriormente proposto, este foi um tópico que mereceu uma análise mais aprofundada.

O primeiro passo dado no âmbito desta investigação foi a verificação da ocorrência do fenómeno de mosaicismo de ploidia em *S. alburnoides*, um fenómeno nunca antes descrito neste complexo híbrido, mas bastante frequente em alopoliploides. A ocorrência de mosaicos de ploidia poderia explicar as diferenças de expressão alélica existentes entre indivíduos triploides e/ou entre órgãos do mesmo indivíduo triploide. Para explorar esta hipótese foram desenvolvidos protocolos de citometria de fluxo e *sorting* de células de forma a identificar indivíduos mosaicos de ploidia e isolar as populações com ploidia diferente permitindo que fossem analisadas de forma independente quanto ao seu padrão de expressão alélica. Apesar da ocorrência do fenómeno de mosaicismo de ploidia em triploides de *S. alburnoides*, esta não explica os padrões de expressão alélica obtidos.

Embora o complexo *S. alburnoides* tenha sido usado como principal modelo de estudo, seguiu-se também uma abordagem multiespécies para comparar a expressão génica e a sua regulação entre peixes diploides e triploides, pois outro objectivo desta tese foi determinar se os mecanismos de regulação da expressão génica previamente especulados para *S. alburnoides* eram uma característica particular deste complexo ou se apresentavam uma ocorrência mais difundida entre vertebrados alopoliploides. Assim, recorrendo a indivíduos diploides e triploides de *Oryzias latipes*, *Poecilia formosa* e *S. alburnoides*, incluindo híbridos tri-genómicos produzidos em laboratório, foram feitas sequenciações de RNA de nova geração, "*de novo assembly*" de transcriptomas, comparação de níveis de expressão génica e de expressão alélica específica e determinação de níveis globais de metilação.

Em relação ao complexo *S. alburnoides*, verificou-se que, apesar dos triploides serem afectados por uma significativa regulação negativa da expressão génica, esta não corresponde a uma "diploidização" funcional do genoma. Em vez disso, para os híbridos triploides, foi sugerida a existência de alguma flexibilidade dos níveis de expressão. Esta tolerância ou flexibilidade pode ser a base da resiliência dos vertebrados inferiores a mudanças de ploidia.

Para além disso, verificou-se também que a regulação negativa da expressão génica nos triploides de *S. alburnoides* não está dependente do silenciamento preferencial de cópias alélicas de um ou de outro complemento genómico, como previamente se havia especulado. Uma análise abrangente da contribuição de cada genoma heteromórfico para os padrões globais de expressão génica revelou a ocorrência de enviesamento da expressão alélica e não uma contribuição equitativa entre alelos. Para um número significativo de genes, foi verificado o favorecimento acentuado da expressão de um ou de outro dos genomas heteromórficos intervenientes no complexo e, em muitos casos, expressão exclusiva a partir de apenas um dos genomas envolvidos (silenciamento alélico). No entanto, a incidência do enviesamento da expressão alélica não foi significativamente afectada pelo nível de ploidia dos indivíduos, e a taxa de silenciamento alélico observada foi semelhante entre diploides e triploides. No decurso desta tese verificou-se ainda a ocorrência de silenciamento em alotriploides de *O. latipes* e de *P. formosa,*

que tal como os triploides de S. alburnoides são resultado de fenómenos de alopoliploidização a cada geração.

Este trabalho permitiu também corroborar a dependência dos padrões de expressão alélica da combinação genómica específica de cada híbrido e também a influência da história evolutiva de cada genoma interveniente. Os resultados apontaram, não só para um efeito notório dos processos evolutivos longos nos padrões de expressão alélica, mas também nos níveis de metilação. Relativamente a esta última, a hipótese de uma regulação negativa da expressão génica mediada pela ocorrência de metilação maciça dos genomas híbridos triploides não foi verificada em *S. alburnoides* nem em *P. formosa*.

Adicionalmente, no âmbito desta tese, foi ainda produzido o primeiro transcriptoma de referência para o complexo *S. alburnoides*.

Em suma, este trabalho demonstra a complexidade da alopoliploidia ao nível da regulação da expressão génica, corroborando a dificuldade em definir regras e/ou explicações universais aplicáveis a todas as condições alopoliploides, salientando a considerável diversidade de mecanismos inerentes a estes organismos.

Palavras-Chave: alopoliploidia, *Oryzias latipes, Poecilia formosa,* regulação da expressão génica, *Squalius alburnoides.*

SUMMARY

Plants, invertebrates and even lower vertebrates are known to deal with hybridization and polyploidy very successfully, surpassing the genetic constrains those phenomena bring. However, (allo)polyploidy in animals have been strongly neglected so, this matter remains largely unexplored. In that sense, the general goal of this thesis was to expand the existing limited knowledge on the topic, standing a significant step forward in the scarce information available on animal allopolyploid gene expression regulation.

The inception of this work was a theory of occurrence of global dosage compensation by allele copy silencing in *Squalius alburnoides* complex. The elucidation of the inherent gene expression processes and mechanisms operating in *S. alburnoides*, and if they are a particular feature of this complex or have a more widespread occurrence among allopolyploids, were the main goals.

The first step taken was the exclusion of ploidy mosaicism, a phenomenon here for the first time described to occur in *S. alburnoides*, as the source of the allele specific expression differences previously found.

Despite it was corroborated that *S. alburnoides* triploids are affected by a significant down regulation of gene expression, that does not correspond to a genome wide exact functional diploidization. Instead, a certain level of flexibility of expression within a range of mRNA amounts per locus was observed. That feature might be a key point in the mechanisms that allow lower vertebrates to endure and maintain ploidy changes so effectively.

The down regulation of gene expression in triploid *S. alburnoides* was also found to be not dependent of allele copy silencing, as previously speculated. Extreme homoeolog expression bias, comprehending the complete silencing of alleles, have been found to affect a significant percentage of genes in *S. alburnoides, as i*n laboratory produced triploid hybrid *Oryzias latipes*. However, the incidence of the homoeolog expression bias was not significantly affected by the ploidy level of the individuals, and the allelic silencing rate was similar between diploids and triploids.

Additionally, the hypothesis of a down regulation of gene expression mediated by massive methylation occurrence in triploid hybrid genomes was not sustained, neither for *S. alburnoides* nor for *P. formosa*.

Keywords: allopolyploidy, gene expression regulation, *Oryzias latipes, Poecilia formosa, Squalius alburnoides*.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1. General background

For many years focus has been put on the negative effects polyploidy and hybridization may have on species evolution. However, in the recent years the emphasis is now on the alternative view, that polyploidy and hybridization are significant creative forces.

It is now established that polyploidy in animals (vertebrates and invertebrates) is not as rare as it was initially assumed. There are in fact plenty examples from all major taxonomic animal groups, except for mammals, of successful polyploidization events (Gregory and Mable, 2005) and is accepted that most vertebrates are ancient polyploids (Wertheim *et al.,* 2013).

It is also long time established the existence of a strong association between polyploidy and hybridization (Mable *et al.,* 2011) so, the wide prevalence of animal polyploidy indicates that natural animal hybridization is also widely prevalent.

Those findings, together with the recent and continuous development of powerful genetic and genomic tools, like next generation sequencing (NGS), have given a new perspective to the general importance and increased interest in understanding the mechanisms and evolution of allopolyploids (polyploids resultant from hybridization events), both in plant and animal kingdoms (Stöck and Lamatsch, 2013). However, research on (allo)polyploids is still highly biased towards plants (Gregory and Mable, 2005). Animal allopolyploid research emerged much later, being significantly delayed most probably due to old rooted ideas that polyploidy is rare among animals and that hybrids are sterile "dead ends".

Because in the past years these ideas have been strongly challenged, the genetic and molecular mechanisms underlying animal allopolyploidy have started to be addressed more enthusiastically (Stöck and Lamatsch, 2013). However,

research on animal allopolyploids it is still far behind from the body of knowledge gathered around plant allopolyploids. (Gregory and Mable, 2005).

1.2. Hybridization and polyploidy

A definition of hybridization is the interbreeding of individuals from two distinct populations or groups of populations, distinguishable by one or more heritable characters (Harrison and Larson, 2014), and comprehending several levels of taxonomical relation (e.g. within same genus, same species or same sub-species).

Hybridization can be homoploid, when occurs between individuals with the same ploidy level and giving rise to hybrids with the same number of chromosomes as the parental species (Yakimowski and Rieseberg, 2014). Homoploid hybridization gives rise mostly to diploid hybrids (Abbott *et al.,* 2013; Yakimowski and Rieseberg, 2014; Nieto Feliner *et al.,* 2017) and stabilized introgressants (Lowe and Abbott, 2015). Homoploid hybrids may be viable and perpetuating, but many present chromosomal incompatibilities and are often sterile due to the impossibility of true chromosomal pairing at meiosis (Stebbins, 1950).

Hybridization can also be associated with changes in ploidy level, either resulting from hybridization between heteroploid entities or resulting from genome duplications after homoploid hybridization (Marques *et al.,* 2017). Polyploidy among hybrid taxa is extremely prevalent and result either from the high rate production of unreduced gametes by diploid hybrids (Ramsey and Schemske, 2002), or as a strategy to overcome the pairing problems between heterologous chromosomes, providing each chromosomal set with a compatible pair (Marques *et al.,* 2017).

Concerning polyploidy, it can be defined as the heritable condition of possessing more than two complete sets of chromosomes (Comai, 2005). It has two fundamental types, depending on the chromosomal composition and formation mechanism. There is autopolyploidy, when the polyploidization process occur isolated, with no hybridization step and involving only homospecific sets of chromosomes (Otto, 2007). On the other hand, as mentioned above, when

complete heterospecific sets of chromosomes are brought together by hybridization (Otto, 2007) there is allopolyploidy.

Autopolyploids usually present polysomic inheritance, multivalent association of chromosomes during meiosis I and no prior differentiation in the chromosomal sets (Stebbins, 1947, Parisod *et al.,* 2010, Wright *et al.,* 2014). In contrast, most allopolyploids are balanced allopolyploids, rarely presenting multivalent association and have the same cytogenetic behavior as diploids since they contain multiples of diploid genotypes in one genome (Stebbins, 1947, Parisod *et al,* 2010, Wright, 2014).

Stebbins (1947) was the first to propose a genetic/cytogenetic approach to distinguish auto-from allo-polyploids, that has been widely used. Nevertheless, it can be deceptive because chromosome pairing can be affected by other factors than the chromosomal structure (Jenczewski and Alix, 2004; Otto, 2007). Nevertheless, polysomic inheritance is considered a good indicative feature to distinguish autopolyploids from allopolyploids and has been regularly observed in natural populations (Jackson and Jackson, 1996; Landergott *et al.,* 2006; Stift, *et al.,* 2008).

Allopolyploid species conceptually have a high potential to adapt to a wider range of ecological niches and to unstable environments, better surviving than their diploid progenitors (Stebbins, 1950, 1971, Mable, 2013). On the other hand, autopolyploids have been found to be rare and suffering from several evolutionary disadvantages compared to their hybrid counterparts (Clausen *et al.,* 1945; Stebbins, 1971). Ramsey and Schemske (1998) estimated that the rate of autopolyploid formation is higher than that of allopolyploids but on the other hand, natural allopolyploidy have been more consensually considered "more common" than autopolyploidy (Abbott *et al.,* 2013).

For now, the relative frequency of autopolyploid versus allopolyploid origins remains uncertain (Doyle and Sherman-Broyles, 2017), and may be difficult to determine as in nature several intermediate situations might occur (Ramsey and Schemske, 1998; Mallet, 2007). In fact, many naturally occurring polyploids may be

incorrectly strictly classified as auto- or allo-polyploids, as they can be intermediates or "segmental allopolyploids" (Stebbins, 1947; Boff and Schifino-Wittmann, 2003).

Through artificially manipulation undoubtable autopolyploids can be produced (Hegarty *et al.,* 2013; Zhou and Gui, 2017) and have a key role in the study of polyploidy. Despite being experimentally less appealing and not the focus of this thesis, the study of autopolyploid systems is in fact also of graet interest, because they are fundamental to disentangle the effects of hybridization and ploidy rise onto gene expression regulation and output (Wang *et al.,* 2006; Chen, 2007).

Further than being auto- or allo-, polyploids can be also classified based on their evolutionary age as neo-, meso- or paleopolyploids, ordered by increasing age (Comai, 2005). Additionally, polyploids occur naturally but can also be the result of experimental breeding (Sattler *et. al.,* 2016).

1.3. The prevalence of animal (allo)polyploidy

It has been accepted since many years that different organisms typically display variable tolerance to polyploidy and hybridization, with plants generally being recognized as more tolerant than animals, and invertebrates more tolerant than vertebrates (Ohno, 1970; Stebbins, 1971). The "why that is so?" or even if "it is that so?" are questions not fully clarified.

The fact is that hybrids and allopolyploids among animals have for long been considered simple errors of the evolutionary process, but the success and perpetuation of hybrid, polyploid and allopolyploid species, as also their significant frequency in many animal groups, like in fish, have been contradicting this old rood ideas.

1.3.1. Historical perspective of "Why polyploidy is rarer in animals than in plants"

Polyploidy has been known and recognized to be particularly prevalent among plants, both as an ancient and an ongoing evolutionary process (Adams and Wendel, 2005). But on the other hand, despite a considerable number of

successful and stable polyploid animal taxa have been revealed over the years, the role of polyploidy in animal evolution has been neglected and the focus has been put on deleterious aspects (Orr, 1990).

The minor role for polyploidy in animal evolution was posed based on the classic arguments presented by Müller in 1925, stating that the animal occurrence of polyploidy would be impaired by its interference with sex determination processes and the disruption of gene dosage balance (Müller, 1925; Orr, 1990). Since then, this paradigm became progressively imprinted, and zoologists significantly reduced their interest in embracing research on the topic. On the contrary, with plants, research on polyploidy experienced a sound development within both conceptual and experimental frameworks. This high disparity between zoological and botanical research is evident in the literature, and a good example of that is the revision edited by Soltis and Soltis (2012) on polyploidy and genome evolution, where only 17% of the compilation being animal-related.

While cytogenetic analysis of large samples of wild specimens is a common practice for botanists, it is much more uncommon for zoologists due to practical and conceptual reasons (Mable, 2004). So, if cytogenetic analysis was further applied to large samples of wild animal specimens, it is expected to greatly increase polyploidy occurrence discoveries within the animal kingdom. In fact, it seems that a positive correlation between records of natural animal polyploids and the efforts putt into taxonomic surveys by animal cytogeneticists can already be noticed (for example in Ráb and Collares-Pereira, 1995; Grishanin *et al.*, 2006; Gromicho and Collares-Pereira, 2007; Hall, 2009).

Interesting is also to note that zoologists have frequently a typological concept of karyotype as in Ráb and Collares-Pereira (1995), Gromicho and Collares-Pereira (2007) and Arai (2011). Also, many have assumed that any structural or numerical change in the karyotype would seriously compromise the fertility of the organism, and consequently will be negatively selected. In that scope, animal polyploids have been seen by zoologists as transitory mistakes with low or none evolutionary weight, and consequently have not been "worthy" of much research efforts. However, for example, in fish, the successful maintenance

and perpetuation of several polyploid species and lineages strongly suggests the opposite assumption (Le Comber and Smith, 2004).

Another aspect to consider is that ancient polyploids are difficult to recognize. One reason is that polyploid genomes can be highly dynamic and post-polyploidization events can be masked by large-scale genome reorganization events that are collectively termed 're-diploidization' (Dodsworth *et. al.,* 2016). Hence, many animals that now appear to be diploid might have been polyploid in origin. Anyhow, it is already established that ancient as well as recent polyploidizations have not only significantly shaped the genomes of plants but also the genomes of animals (Gregory and Mable, 2005). In summary, it is not by far a marginal process in animal evolution.

1.3.2. Polyploidy in animals

Concerning invertebrates, polyploidy has been documented nearly in all major phyla (Turbellaria, Trematoda, Nematomorpha, Tardigrada, Aracnidae, Rotifera, Insecta, Crustacea, Annelida and Mollusca) (Gregory and Mable, 2005).

Among vertebrates, polyploidy is especially frequent among fish, but it is also frequent in amphibians and reptiles (Mable *et al.,* 2011). The classical examples of polyploid amphibians are the anuran species *Hyla versicolor* (the Gray treefrog), *Phyllomedusa burmeisteri* (Walking leaf frog) and *Odontophrynus americanus* (the American ground frog) (White, 1978; Otto and Whitton, 2000) but other examples can be found in Urodela, as for example the *Ambystoma jeffersonianum* (Jefferson salamander) (Otto and Whitton, 2000). In reptiles, mostly in lizards and in one species of turtle, the *Platemys platycephala* (twist-necked turtle), polyploidy was also documented (Otto and Whitton, 2000; Gregory and Mable, 2005; Bogart *et. al.,* 2007; Mable *et al.,* 2011; Evans *et al.,* 2012).

In fishes the abundance of polyploidy examples in four highly species rich families (Acipenseridae, Catostomidae, Salmonidae, and Cyprinidae) (Otto and Whitton, 2000) have granted them a title of exemption regarding the generalized assumption of rarity of polyploidy in vertebrates. In fact, there are already several extensive studies that were dedicated to animal polyploidy which

showed/postulated a preponderant incidence in fish, for example, Leggatt and Iwama, (2003); Le Comber and Smith (2004) and Mable *et al.,* (2011). To date, polyploidy in fish is hypothesized for Petromyzontiformes and is now confirmed in the classes Chondrichthyes, Sarcopterygii and Actinopterygii, in a total of 14 orders (Collares-Pereira *et al.,* 2013).

Among homeotherm vertebrates, polyploidy was found less widespread. It was erratically documented in *Gallus domesticus* (domestic chicken) and in a Psittaciform species, the *Ara ararauna* (blue-and yellow macaw) (Otto and Whitton, 2000; Gregory and Mable, 2005). It was also described for two octodontid rodents from Argentina's deserts known as the Red and the Golden Viscacha rats (*Tympanoctomys barrerae* and *Pipanacoctomys aureus*) (Gallardo *et. al.,* 1999, 2004, 2006), but those findings have been strongly disputed (Svartman *et al*., 2005). In humans, constitutive total polyploidy also occurs, either in the form of triploidy or tetraploidy. It is estimated to happen in every 4.5–8.8% of human conceptions, but it leads consistently to an early developmental inviability (Egozcue *et al*., 2002).

1.4. Fish as good models to study (allo)polyploid genome regulation.

Polyploidy has been discussed as being particularly important and prevalent with regard to speciation of fishes (Le Comber and Smith; 2004). In fact, fishes are great model systems to study the origins and consequences of polyploidy. They offer opportunities to study and understand polyploidy in contexts, which are harder to find in plants. For example, dioecy is much rarer in plants than in animals (Renner and Ricklefs, 1995; Bull, 1983), and it is frequently genetically based. In genetically based sex determination systems, issues may emerge with ploidy increase (e.g. correct gene dosage achievement) (Müller, 1925; Mable, 2004). However, those issues have been creatively and diversely overcome in many polyploid fishes. Despite the variability of sex determination processes has been demonstrated among different vertebrate taxa (Grandont and Jenczewski, 2013; Stenberg and Saura, 2013), no other group offers so much diversity of sex determination mechanisms as teleost fish. In teleosts, a complete set of reproductive systems can be observed and studied in parallel with polyploidy.

Strictly bisexual and unisexual polyploid fish do occur (Tsigenopoulos *et. al*., 2002), but also many other types. These include asexual reproduction, hermaphroditism, sex determination based on the presence of a single or multiple central regulatory sex chromosome or based on the ratio of sex chromosomes to autosomal chromosomes (Otto and Whitton, 2000; Mank and Avise, 2009; Matos *et al.,* 2010; Machado *et al.,* 2016). All are common sex determination mechanisms found in fishes, strictly, simultaneously or sequentially. Additionally, because of the frequent convergence of hybridization with polyploidy, many altered distinct reproductive pathways with altered oogenesis and/or spermatogenesis (eg. meiotic hybridogenesis, clonal gametogenesis, non-reductional meiosis) are used by allopolyploid fish (Alves *et al.,* 2001).

Another asset of fishes that increases their value as model system in which to examine polyploidy is that polyploidy can be studied at different time scales. The ancestors of vertebrates went through two genome doubling events, one prior to the Cambrian explosion, and a second one in the early Devonian (Meyer and Schartl, 1999). Then, in the late Devonian, after the divergence of the lobe-finned fishes (Sarcopterygii) there was a third duplication, leading to the ray-finned fishes (Actinopterygii) (Amores *et al.,* 1998; Vogel, 1998; Meyer and Schartl, 1999). As a consequence of this duplications, the multigene families in fish are larger than in mammals (Meyer and Schartl, 1999) and the resultant increased redundancy in gene copy number have been pointed out as a possible cause for the success and diversity of fishes (Meyer and Schartl, 1999).

1.4.1. The hybridogenetic fish complex *S. alburnoides*

S. alburnoides is described as an allopolyploid hybridogenetic complex, endemic from the Iberian Peninsula. "Hybridogenetic" refers to an alternative mode of reproduction resembling parthenogenesis, but rather than completely asexual, it is hemiclonal (Alves *et al*., 2001). "Complex" is the technical terminus denoting a natural system composed of parental species and their hybrids, with altered modes of reproduction and reproductive interdependence (Alves *et al*., 2001).

In the origin of the *S. alburnoides* complex is a unidirectional hybridization event involving the sympatric *S. pyrenaicus* females (P genome) (Alves et al., 1997) and males from an already extinct species belonging to the *Anaecypris hispanica* lineage (A genome) (Alves *et al.,* 2001; Robalo et al., 2006; Gromicho *et al.,* 2006a; Collares-Pereira and Coelho, 2010). Concerning the age of the complex, as also concerning the number of original hybridization events, there is some controversy. While Cunha *et al.* (2004) suggested an origin around 1,400,000 years ago, Sousa-Santos *et al*. (2007a) estimated an age of 700,000 years. In terms of reproductive modes, both sexual and nonsexual are found within this complex and the intervenient individuals are neither strictly clonal nor hemiclonal regarding their inheritance patterns (Crespo-López *et al*., 2006; Gromicho and Collares-Pereira, 2007).

The complex is widely distributed in the Iberian Peninsula (Figure 1).

Figure 1. Iberian Peninsula - Distribution range of *Squalius alburnoides* and of the 3 *Squalius* species which contribute to the reproductive dynamics of the complex.

From the mate of individuals with the original hybrid genomotype (PA) with individuals of other sympatric *Squalius* species (*S. carolitertii* in the Northern basins – C genome; *S. pyrenaicus* in the southern basins – P genome and *S. aradensis* in the independent southerly Quarteira basin – Q genome), several allopolyploid forms emerged, including forms with a replacement of the *S. pyrenaicus* nuclear ancestral genome by the genomes of the extant sympatric species through introgression (Alves *et al*., 2001; Pala and Coelho, 2005; Sousa-Santos *et al*., 2006a; Cunha *et. al*., 2008).

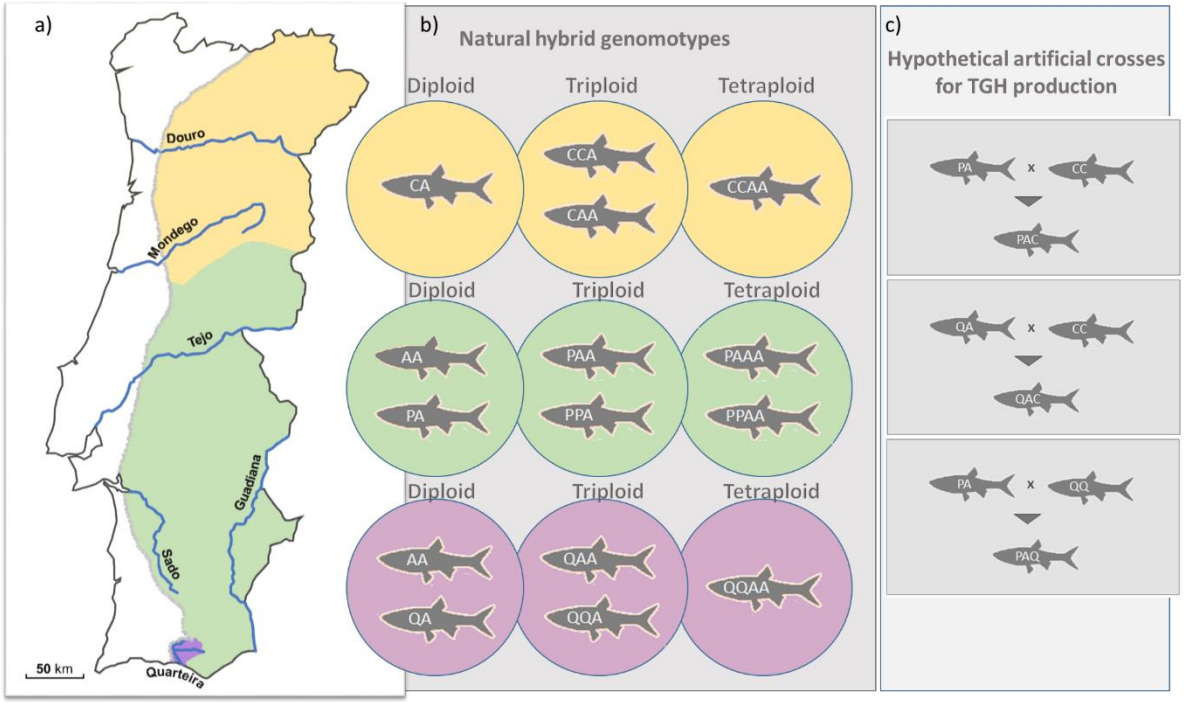


Figure 2. Different *S. alburnoides* genomotypes. a) hybrid zones; b) genomotypes found in natural populations; c) hypothetical artificial crosses leading to tri-genomic hybrid (TGH) progeny.

Nowadays, in Portugal, the complex comprise fertile and interdependent hybrids of both sexes, with several ploidy levels and genomic compositions including diploids (2n=50: PA, CA; QA;), triploids (3n = 75: PAA, PPA, CAA, CCA; QAA; QQA) and tetraploids (4n = 100: PAAA; PPAA, CCAA; QQAA) (Figure 2b)(Alves *et al*., 2001; Collares-Pereira and Coelho, 2010; Collares-Pereira *et al.,* 2013) (Figures 2a and 2b), and despite not yet found in nature, other triploid and

tetraploid genomic combinations can be obtained in captivity. For example, triploids with three different genomic complements (tri-genomic hybrids) can be produced, either with PAQ, PAC or CAQ combinations (Figure 2c).

The *S. alburnoides* mtDNA is usually *S. pyrenaicus* (the maternal ancestor), although some other introgressions have been reported (Alves *et al*., 1997; Sousa-Santos *et al*., 2006a, 2007a).

There are significant differences between populations concerning the relative frequency of each genomotype of *S. alburnoides* (Collares-Pereira *et al*., 2013). However, in common they have a substantial sex bias towards females. In fact, triploids are mostly females, and triploids are the most abundant form in almost all populations of *S. alburnoides*. The exceptions are a couple of populations found in the Northern Portuguese basins, which are mainly constituted by symmetrical tetraploids (CCAA) with a balanced sex ratio (Cunha *et al*., 2008).

A major difference between populations is the exclusive presence in central and southern populations of an all-male lineage of nuclear non-hybrids with AA nuclear genome and *S. pyrenaicus* (P) mtDNA. All AA individuals found so far were males, with exception from 2 females, 1 found by Carmona *et al*. (1997) and other by Sousa-Santos *et al*. (2006b). These males do not represent neither an independent nor a self-sustainable lineage as they are solely reconstituted within the complex by mating with allotriploid females (PAA or QAA).

Concerning the reproductive modes of *S. alburnoides*, they are highly diverse. Normal sexual reproduction, clonal inheritance, hybridogenesis and meiotic hybridogenesis have long been identified to occur in *S. alburnoides* (reviewed in Alves *et al*., 2001 and Collares-Pereira *et al*., 2013), but recently also androgenesis was added to this list (Morgado-Santos *et al*., 2017). Moreover, the reproductive interdependence between forms is highly complex, but it is considerably well-known for the most abundant forms of the complex (Alves *et al*., 2001; Collares-Pereira *et al*., 2013; Morgado-Santos *et al*, 2016; 2017), and grounded on artificial crosses and analysis of distinct molecular markers (Alves *et al*., 2001; Pala and Coelho, 2005; Crespo-López *et al*., 2006; Morgado-Santos *et al*., 2016; 2017). On the other hand, the reproductive modes of rarer genomotypes, such as

unbalanced tetraploids and most the genomotypes containing *S. aradensis* genome (Q), have not yet been identified.

Despite the diversity of reproductive modes found in *S. alburnoides*, the most characteristic of the complex, and the one undertaken by the most common *S. alburnoides* genomotypes (PAA), is meiotic hybridogenesis. In meiotic hybridogenesis the heterospecific genome is discard from the oocytes and the remaining (similar) genomes undergo a normal meiosis (Alves *et al*., 2001; Sousa-Santos *et al*., 2007b).

Although *S. alburnoides* individuals have distinct reproductive modes, they are gonochoristic, but a very few hermaphroditic individuals have been identified (Matos *et al*., 2010). Also, *S. alburnoides* lacks sexual dimorphism so, the sex of the individuals can only be determinable inside the reproductive season (once a year from March to May). Other interesting feature of the *S. alburnoides* complex is that a visual distinction between genomotypes is not a simple thing. For example, diploid and triploid hybrid forms are undistinguishable by morphometric characters (Cunha *et al*., 2009). On the other hand, PP, CC and AA genomotypes are easily distinguishable from each other and from the hybrids. Also, very few differences in growth and reproductive traits were found between PA and PAA females (Ribeiro *et al.*, 2003). Concerning longevity, also only marginal differences have been observed between triploid (living up to 6 years), diploid females (living up to 5 years) and AA males (living up to 4 years) (Riberio *et al.,* 2003).

Due to its unique features among polyploid taxa, the *Squalius alburnoides* complex of hybrid fish has been a desirable system to study genome regulation and interaction in animal allopolyploids.

As an inter-generic allopolyploid, sequence differences have been easily found and used to discriminate between different genome-specific gene copies and determine if they contribute or not to the overall expression (Pala *et al,* 2008a; 2010). Coupled with the diversity of ploidy levels and genomic constitutions, the complex offers a multitude of hybridization and ploidy scenarios to be studied. As a result, it was the first allopolyploid vertebrate model established to address questions on gene expression regulation (Pala *et al*., 2008a) and genomic

interactions (Pala *et al*., 2010). Furthermore, the presence of lineages within the complex, established differentially in time, further allowed to start exploring the evolutionary perspective of the mechanisms of gene expression regulation in vertebrate allopolyploids (Pala *et al*., 2010). However, even with a promising starting point of genetic information provided by Pala *et al*., (2005; 2008a; 2008b; 2010) a wide-scale analysis of gene behavior throughout whole genomes have not yet been achieved due to lack of high throughput sequence data at that time.

1.4.2. The gynogenetic fish complex *P. formosa*

The Amazon molly (*Poecilia formosa*) is a small fresh water, live bearing fish that occurs in the Atlantic drainages of Central America, from Rio Tuxpan, Mexico, to South Texas, U.S.A. (Figure 3).

Figure 3. Mexico Gulf - Distribution range of *P. formosa*, *P. mexicana* and *P. latipinna* species.

The species is named after the Amazons, the mythical all-female tribe of warriors that used males from neighboring tribes to mate, and by killing all the resultant male progeny perpetuated as an all-female group. As the mythic Greek Amazons, Amazon mollies are all females, but instead of killing their male offspring, they simply do not produce them.

As an all-female species *Poecilia formosa*, stands out as a vertebrate aquatic model. They present genetic clonality, a direct result of its unusual mode of reproduction, gynogenesis, which is a sperm-dependent parthenogenesis (Lampert and Schartl, 2008). *P. formosa* produce diploid eggs in absence of meiosis and these diploid eggs are pseudo-fertilized by sperm of males of closely related gonochoristic (bisexual) species (Lampert and Schartl, 2008). So, in nature, Amazon mollies always coexist with at least one of these three species: the Sailfin molly (*Poecilia latipinna*) in the USA and northern Mexico, the Atlantic molly (*P. mexicana*) and the Tamesi molly (*P. latipunctata*) in Mexico. These species are known to serve as sperm donors in the *P. formosa* natural habitats (Lampert and Schartl, 2008). However, the sperm normally do not contribute genetic information, being only used to trigger embryogenesis of the eggs (Lampert and Schartl, 2008). The paternal pronucleus does not fuse with the unreduced diploid oocyte nucleus, and the paternal genetic material is expelled. So, the vast majority of *P. formosa* are diploid and genetically identical (clones) to their mothers. But, in some rare instances paternal introgression occurs (Lampert and Schartl, 2008). Either small parts of male genetic material are included as microchromosomes (Nanda *et al*, 2007), or the sperm nucleus may indeed fuse with the oocyte nucleus resulting in triploid offspring. Such triploids found in the wild are fertile and produce clonal all triploid female offspring (Lampert et al., 2005; 2007). However, the allopolyploidization events as the origin of such individuals are extremely rare and are considered ancient occurrences as they were traced back to the evolutionary past of *P. formosa,* in two instances (Lampert *et al*. 2005; Schories *et al*. 2007).

In laboratory broods, allotriploids have been also obtained from diploid *P. formosa* as rare introgression cases of paternal genomes of closely related species (Nanda *et al*. 1995). Besides *P. latipinna, P. mexicana* and *P. latipunctata* that were

previously mentioned, also aquarium/ornamental strains like Liberty mollies (derived from *P. salvatoris*) and Black mollies, are commonly used for that purpose. On the contrary to their natural triploid counterparts, this laboratory produced *de-novo* triploids do not give rise to stable gynogenetic lines and can present different genomotypes depending on the parental species used for breeding. These include three different genome hybrids *P. formosa*. Different tri-genomic hybrids (TGHs) can be produced, for example, from *P. formosa* diploids (with *ml genome)* with introgressed genome from *P. salvatoris*, (*s genome*), or *P. formosa* diploids (*ml* genome) with introgressed genome from black molly (*b genome*) (Lamatsch *et al.,* 2010). *TGHs* are promising models for studying allele specific expression (Figure 4).

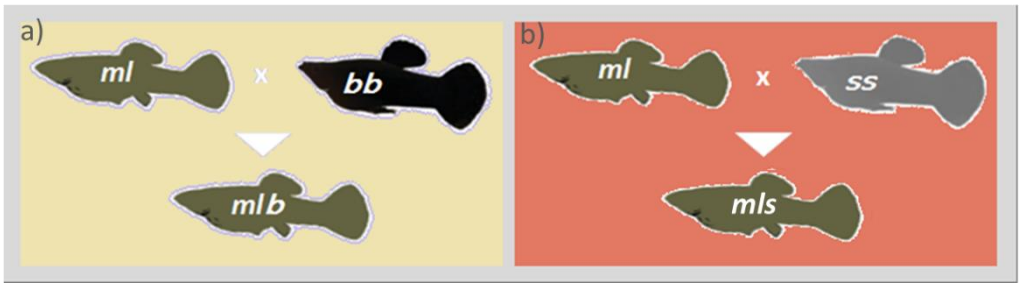


Figure 4. Laboratory crosses leading to tri-genomic hybrid (TGH) progeny. a) *P. formosa* diploids (with *ml genome)*, with introgressed genome from black molly (*b genome*) and b) *P. formosa* diploids (*ml* genome) with introgressed genome from *P. salvatoris*, (*s genome*).

Poecilia formosa has also been used as disease model, for example in cancer studies (Schartl *et al.,* 1997; Woodhead *et al.,* 1984) and infectious diseases (Tobler and Schlupp, 2005), but above all, the *Poecilia formosa* complex has been so far, an emblematic model in evolutionary biology (Tobler and Schlupp, 2005), mostly concerning the costs and benefits of sexual vs asexual reproduction (Schlupp, 2010).

Despite that there is still a huge scarcity of genomic resources for *Poeciliids* in comparison to more conventional fish model organisms as Zebrafish, the Amazon molly genome sequence (GenBank Genome ID: 13072) and the genomes of *P. latipinna* (GenBank Genome ID: 17477) and *P. mexicana* (GenBank Genome ID: 14658) have been recently published (Warren *et al.,* 2018).

1.4.3. *The convenient laboratory engineered model Oryzias latipes*

Medaka (*Oryzias latipes*), which in Japanese means "tiny fish with big eyes", is a small, egg-laying freshwater fish that since more than a century has been helping to pave the scientific path (Shima and Mitani, 2004; Naruse *et al*., 2011). It was scientifically first described as *Poecilia latipes* in 1850 and later renamed in 1906, as *Oryzias latipes*. The name reflects the preference of Medaka (ricefish) to live in the rice (*Oryza sativa*) fields (Wittbrodt *et al*., 2002).

In captivity, under laboratory-controlled conditions of 14h light/10h dark, temperature between 25º to 28ºC and successive mattings, the live span of Medaka is about 1 year. If they are not allowed to mate, put under a light/dark cycle of 10h light/14h dark and at a lower temperature of around 19ºC, Medaka can live about 2 years (Kirchmaier *et. al*., 2015).

Medaka males and females present obvious sexual dimorphism. A slit in the dorsal fin and the hooks on the anal fin rays are evident in the males.

Medaka is native to Taiwan, Korea, China and Japan (Shima and Mitani, 2004). Genetic differentiation among populations has been showed by phylogenetic analysis (Sakaizumi *et. al.,* 1983 Matsuda *et al*., 1997; Takehana *et al*., 2003, 2004, 2005), and the Medaka populations have been initially classified into 4 genetically divergent groups: The Northern Japanese, the Southern Japanese, the Eastern Korean and the China-Western Korean (Takeda and Shimada, 2010) (Figure 5).

Figure 5: Sea of Japan - Distribution range of the four major Medaka groups in East Asia: China and Western Korean, Northern Japanese, Southern Japanese and East Korean. From the natural populations, laboratory inbred lines HNI, Hd-rR, and HSOK have been derived.

Previously it was thought that the southern and northern Japanese populations of Medaka were the same species (*O. latipes*), but recently the northern population has been classified as a new species, *O. sakaizumii*. O. sakaizumii is now known as the northern medaka and *O. latipes* as the southern medaka (Matsuda and Sakaizumi, 2015).

The haploid genome size of Medaka in approximately of 800 Mb (Kasahara *et al*., 2007) and the diploid karyotype of Japanese and East Korea populations is of 48 chromosomes, while medaka from West Korea and China have 46 chromosomes due to a fusion of chromosome 11 and 13 (Uwa and Ojima, 1981).

Medaka is highly tolerant to inbreeding and this feature has been exploited decades ago to establish highly inbred strains from different wild populations (Hyodo-Taguchi 1980). As in some cases these strains have been inbred for more than 100 generations they can be considered as isogenic.

The northern and southern groups diverged approximately 4 to 18 million years ago. Presently, the degree of nucleotide polymorphism between inbred strains of these groups is extremely high, for example between HNI, from the Northern group and Hd-rR from the southern group it is in the range of 1% in coding and 4% in noncoding regions. Northern and southern groups are highly polymorphic and isogenic inbred lines have been established from both groups (Naruse *et al*. 2004). Besides polymorphism between laboratory Medaka strains, there are also marked behavioral differences, body shape differences, strain specific brain morphology and strain specific susceptibility to mutagens (Ishikawa *et al*. 1999; Kimura *et al*. 2007).

Despite the zebrafish is still by far the fish model of election of the scientific community, probably due to its excellent biological features, abundant data and molecular resources, other fish species like medaka are being identified as good or better suited for answering several questions (Schartl, 2014). Although still much less used and with less information available, the medaka is equivalent in many ways to Zebrafish as a model system. For example, both have similar size, short generation time, are easy to breed, to maintain and to genetically manipulate under

laboratory conditions (Schartl, 2014). In fact, concerning genetic manipulation, genome editing with transgenic technologies as TALEN and CRISPR/Cas9 have been successfully and usefully applied to medaka (Zhu and Ge, 2018). There is also increasingly growing data availability and molecular resources on medaka. Several examples of presently available genomic information on medaka are reviewed in Kirchmaier *et al*., (2015). For instance, 1) the medaka genome is already sequenced and can be freely accessed through several genome browsers; 2) the medaka reference genome is based on the Hd-rRII1 inbred line but the genome sequence of HNI-II, Kaga, HSOK and Nilan and Kiyosu strains (Spivakov *et al*., 2014) are now also available; 3) blast searches against Hd-rR II and HNIII scaffolds as well as against raw shotgun reads of Hd-rRII are possible to do; and 4) searches for SNPs in the HdrR, HNI, Nilan, HSOK, and Kaga strains can also be performed.

Using medaka strains to produce synthetic allopolyploids have also been done (Figure 6) (Wakamatsu, 2008).

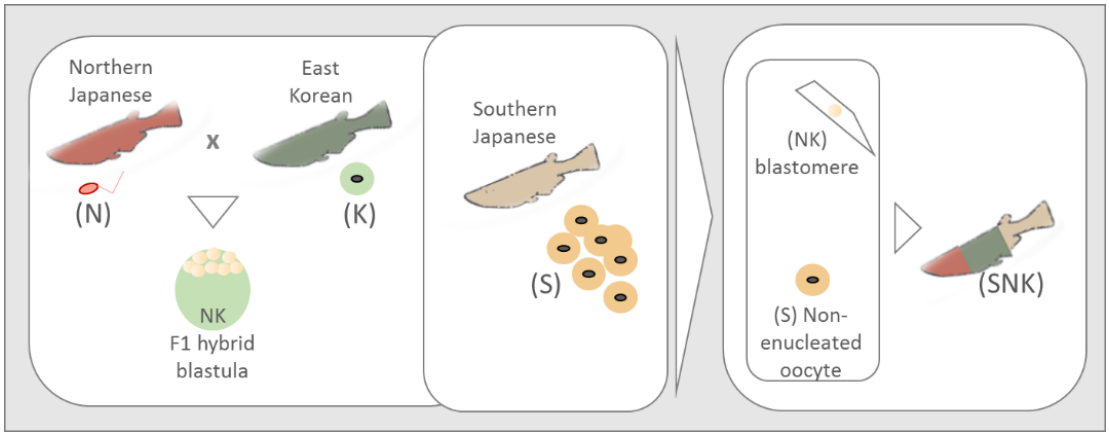


Figure 6. Allotripoid medaka produced by nuclear transfer. Using three different strains, HNI-II (N), Orange-Red (S), and SOK (K) strains, each of which originated from different natural populations (North Japan, South Japan and East Korean, respectively). The nuclear transfer technique consisted in obtaining donor cell nuclei from blastula embryos of F1 hybrids between HNI-II and SOK, and on its transfer to recipient unfertilized eggs of OR, producing triploid tri-hybrid fish.

These synthetic medakas are a very convenient model system for studies in allopolyploidy, with the advantages of having plenty of data bases and molecular

resources to study and to characterize radical genomic changes at early evolutionary states.

1.5. Genome regulation and interactions in animal allopolyploids.

1.5.1. Allopolyploid genome puzzling questions

The allopolyploidization process, namely, the addition of one or multiple complete sets of chromosomes inherited from different strains, sub-species, species, etc., results in increased total DNA content, increased number of alleles at each locus, increased heterozygosity and potentially increased interactions among loci (Johnson *et al*., 2007). These fundamental changes modify relationships within and between loci, with resulting alterations in gene expression and phenotype (Chen, 2007). Hence, allopolyploidization is one of the most dramatic changes a genome can endure. So, it is absolutely fascinating how this phenomenon is so well tolerated and common in nature.

(Allo)Tripoloidy is the polyploid state or condition of having three complete sets of chromosomes. Remarkably, even this euploid/uneven chromosomal state does not always result in catastrophic genomic imbalance. In mammals, it has serious deleterious effects (Antonarakis *et al*., 2004), but lower vertebrates, like reptiles, amphibians and fish cope very well with it (Otto and Whitton, 2000). This implies that either effective compensation is occurring or that there is no necessity for such mechanisms.

It is important to understand what factors allow lower vertebrates to endure and maintain ploidy changes (Pignatta and Comai, 2009). Not only to know for the sake of knowledge but also having in perspective that it may shed some light on how to overcome deleterious expression of supplementary number of chromosomes that occur in higher vertebrates, and generally have very undesirable consequences in humans, like Down Syndrome (Antonarakis *et al*., 2004) and many cancers (DePamphilis, 2016).

1.5.2. Gene expression retort to dosage increase and the pioneer model *S. alburnoides*.

Lower vertebrates deal with hybrid constitution and gene dosage increase very effectively, as they survive and perpetuate (Otto and Whitton, 2000). This evolutionary success suggests that they developed, and/or make use of mechanisms that allow to overcome genomic shock and instability, potentiating species adaptation and viability. The effect of dosage on gene expression is the result of the stoichiometric interactions of multiple dosage sensitive trans-regulatory factors among themselves and with their target genes (Birchler and Veitia, 2007; Malone *et al*., 2012). In euploids the dosage of the genomic complement is changed proportionally, and the stoichiometric relationships are presumably maintained (Birchler and Veitia, 2010). Hence, in theory, the level of gene expression should be increased or decreased according to the ploidy variation, which in many cases does not happen (Birchler *et al*., 2010). A proposed regulatory mechanism to be operating in allopolyploid organisms, that could allow an efficient competition with their diploid equivalents, is dosage compensation (Pala *et al*., 2008a).

As no obvious phenotypic differences are observed between diploid and triploid hybrid biotypes of *S. alburnoides* (Alves *et al*., 2001; Cunha *et al*., 2009) the occurrence of dosage compensation in triploids to the diploid level was easily speculated. The validation of this hypothesis at the expression level was first pursued by Pala *et al*. (2008a).

First, from simultaneously extracted liver DNA and RNA of 2n PA and 3n PAA individuals, the ratio between β-actin transcripts to β-actin gene dosage was estimated. Although the numerical quantification was not easy, due to material amount limitations, a robust tendency towards lower RNA:DNA ratio in triploids was observed. Then quantitative real-time RT-PCR analysis of muscle, eye, liver, and gonad transcripts of these diploid and triploid fish was also performed. Three housekeeping genes (*β-actin, rpl8, and gapdh*) and three gonad-specific genes (*amh, dmrt1, and vasa*) were analysed. No significant expression level differences were found between diploid and triploid samples. This was taken as strong

indication that in triploid fish there was a reduction of gene expression to the diploid level, and so, dosage compensation was operating in here. Later (Pala *et al*., 2010), for the genes *b-actin, rpl8, gapdh* and *ef1a*, relative expression ratios were obtained from the comparison between average real time RT-PCR ct values of triploid and tetraploid samples of several genomic compositions (PAA; CAA; CCA; CCAA) and the diploid controls (PA; CA). It was observed that the 3n/2n and 4n/2n ratios were always approximately 1, which implies dosage compensation by regulation of gene expression to the diploid level in different ploidy levels and genomic compositions.

1.5.3. Dosage compensation by gene copy silencing in the *S. alburnoides* complex

Having sustained the hypothesis of dosage compensation in the *S. alburnoides* complex, the next logical question was "how does it work". The most evident possibility was the transcriptional silencing of a whole genome in triploids. To assess this possibility (Pala *et al*., 2008a, 2010) the allelic expression patterns of four ubiquitously expressed genes (β-*actin, rpl8, ef1a* and *gapdh*), two gonad specific genes (*amh* and *dmrt1*) and one eye-specific gene (*rhodopsin*) were examined in diploid, triploid and tetraploid specimens.

To follow the expression of different alleles, RFLP's and Sanger sequencing of these gene transcripts were performed, and gene by gene, specific expression patterns were determined by the presence or absence of expression of the heteromorphic alleles (P and A, and C and A). Exclusive contribution of alleles of the A genome in some organs and/or genes of triploid PAA and CAA individuals was found, indicating that the P or C genome alleles would be inactivated in these samples. On the contrary, exclusive expression of the P or C genomes (unpaired minority represented genome) in triploid PAA's and CAA's was not detected (Pala *et al*., 2008a and 2010). Therefore, preferential usage of A alleles in PAA's could be interpreted as matter of genomic homology to be at play in regulating the profiles of allelic expression of triploid individuals.

All these findings agree with the hypothesis of dosage compensation by silencing of only one allele in the triploids of *S. alburnoides*. It was proposed that the observed dosage effects on the allotriploids could be the result of the expression of the two homomorphic A alleles when P or C allele is not detected, and the result of silencing of one of the A copies, when both heteromorphic alleles are contributing to the overall expression of a gene. With these results, an apparently parsimonious hypothesis, of whole haplome inactivation (Auger *et al*., 2005) was excluded. Also, parental determined genomic imprinting (Alleman and Doctor, 2000) does not fit, because organ-specific differences were found. Another option, random inactivation, was also discarded because the expected 1:2 ratio of AA to PA and/or AA to CC allelic expression per gene in triploid samples was not respected. An overall predominance of P genome-copy silencing for the analyzed genes was calculated for PAA samples (Pala *et al*., 2008a) and an almost exclusive biallelic expression (CA) in C genome containing triploids was observed (Pala *et al*., 2010).

1.5.4. Genomic context driving the patterns of allelic expression in *S. alburnoides*

Pala *et al*. (2010), using the roughly geographical location north vs south, revealed by the presence of C or P genomes in the allopolyploid biotypes, exposed a substantial difference in genome specific allele usage between genomic contexts. A preferential expression of A genome and silencing of P genome alleles was observed in most triploids of one southern population analyzed (Sorraia river, Tejo basin). Conversely, in two analyzed northern populations (Douro and Mondego river basins), in the vast majority of the samples, simultaneous expression of both C and A genome alleles was detected, irrespective of ploidy level or genomic composition. As such, the different patterns of allele specific silencing found within the complex are apparently dependent on the presence of P or C genome in the triploid hybrids.

It is well documented, mostly in plants (Adams *et al.,* 2003; Rapp *et al.,* 2009; Collares-Pereira *et al*., 2013) but as well in invertebrates (McManus *et al*., 2010) that

the reunion of diverged regulatory systems in a hybrid organism produces different patterns of target gene expression. In *S. alburnoides*, despite P and C genomes having a good functional affinity with the A genome (validated by the viability and apparent equal success of both hybrid biotypes within each specific distribution area), the results from Pala *et al*. (2010) indicate a differential activity of the regulatory elements in the presence of P or C genomic complements.

It is also known that the regulation of gene expression is different between ancient and newly formed polyploids (Adams *et al*., 2004; Adams and Wendel, 2005) which renders the origin and timing of the polyploidyzation and/or hybridization occurrence relevant for the final expression outcome. In this context, *S. alburnoides* complex emerged around 1.4 MY ago according to Cunha *et al*. (2004) and less than 0.7 MY according to Sousa-Santos *et al*. (2007a). Yet, the colonization of the northern basins in Portugal and the introduction of the C genome through *S. carolitertii* is a much more recent event – 0.05 MY ago from Tejo to Mondego and only 0.01 MY ago from Tejo to Douro (Sousa-Santos *et al*., 2007a). So, the later acquisition of the C genome can be the cause for the different interaction with the A genome in the northern allotriploid biotypes.

The differential patterns of gene expression according to genomic composition, for the *S. alburnoides* complex, point towards a strong influence of the type of genomes involved in the hybridization events occurring in each local population.

1.5.5. Candidate regulators of gene expression in allopolyploid S. alburnoides.

The direct players of gene expression, determining which genes are actively expressed and which remain silent, are the transcription factors. Transcription factors are proteins that recognize and bind to specific sequences of nucleotides enabling the assembly and action of the full transcriptional machinery upon the gene body sequences they are regulating. Nevertheless, other levels of complexity of the genome expression regulation than the simple availability of transcription factors have been disclosed over the years (*Blighe et al.,* 2018). Epigenetic marking

and miRNAs condition the genome response to transcription factors and so, shape the gene expression programs in all cells. In the hybrid and polyploid contexts their role and mechanism of action has been also explored (Li *et al.,* 2011; Greaves *et al.,* 2015; Jackson, 2017), and as in any cell, both epigenetic marking and miRNAs are used by the (allo)polyploid cells to modulate gene expression.

Concerning miRNAs, from genome-wide expression studies in allopolyploid plants, their involvement in hybrid and polyploid regulation was exposed (Hegarty *et al*., 2006; Ha *et al*., 2009). Several observations indicate that many genes and miRNAs are expressed non-additively (after hybridization events and/or ploidy increase). Later, for animals, an inverse correlation was established between miRNAs levels and the abundance of transcripts containing complementary binding sites for that specific miRNAs (Lim *et al*., 2005). As it is now known that animal miRNAs can induce target mRNA degradation and the molecular mechanistics behind it (Huntzinger and Izaurralde, 2011), Inácio *et al*. (2012) considered them as good candidate regulators for the observed silencing and compensation in *S. alburnoides*.

With high-throughput arrays and sequencing technologies, and using Zebrafish as reference, the small RNA profiles in different genomic compositions interacting in the *S. alburnoides* complex (AA; PP; PA and PAA) were assessed and compared (Inácio *et al*., 2012). It was verified that diploid and triploid hybrids shared most of their small RNA sequences, and that the miRNA expression profiles between libraries were highly correlated. Yet, an overall view indicates an up-regulation of several miRNAs in triploids and a global miRNA expression in triploids higher than the predicted from an additive model (Inácio *et al*., 2012). The results of this study significantly support that miRNAs are probably promoting or deeply involved in the genome stability that consents the evolutive success of the complex.

Concerning to epigenetics, it refers to heritable properties of the genome not involving alterations of the DNA sequence, and are mediated by chromatin state (Bird, 2007). Epigenetic regulation of gene expression occurs by DNA and/or histone modifications and is widely associated with several phenomena including

gene silencing (Li et al., 2011). So far, the molecular mechanism responsible for allelic silencing and gene expression down regulation in the allotriploid *S. alburnoides* is unknown, but a reasonable explanation, so far not explored in this complex, is epigenetic regulation.

It is a fact very well documented in plants, that hybridization and polyploidy events are often accompanied by epigenetic alterations. As epigenetic changes can be stable but also potentially reversible, epigenetics has been proposed as an effective and flexible mean to face the genomic shock and pass through the incompatibilities (allo)polyploidy may bring (Chen, 2007).

Also, as the post-(allo)polyploidization genome evolution scenarios include restoring of a diploid-like state (Comai, 2005; Zhou and Gui, 2017), it is easy to think that gene silencing events epigenetically mediated and initially reversible, can be in time converted in genetically fixed and irreversible states.

In animals the general topic of epigenetic changes associated to ploidy rise and hybridization has been only barely tackled, with very few examples of studies found on the topic, and all focusing on DNA methylation changes (Koroma *et al.,* 2011; Xiao *et al.,* 2013; Covelo-Soto and Leunda, 2015; Jiang *et al.,* 2016; Zhou *et al.,* 2016; Zhu and Gui, 2017). Concerning histone modifications specifically in (allo)polyploid animals no references were found, but the structural role of histones and the proteins themselves, are highly conserved evolutionarily (Over *et al.,* 2014) and in addition to their structural role, histones can also influence gene expression (Yadav *et al.,* 2018). Histones tails, that are exposed in nuclear environment can be chemically marked, altering the affinity between them and the DNA. The result is a local alteration of the chromatin packing and consequently accessibility of the transcriptional machinery to that DNA area (Over *et al.,* 2014). For example, acetylation of histone tails by histone acetyltransferases loosens the contact with DNA. That creates binding sites for transcription factors enabling or facilitating gene expression. Methylation of histone tails can be either activating or repressive, it depends on the specific amino acid where it occurs (Greaves *et al.,* 2015).

DNA methylation is a universal epigenetic phenomenon (Li *et al*., 2011) and is the most important epigenetic change found to be associated with plant

polyploidy and hybridization (Li *et al.,* 2011; Greaves *et al.,* 2015). Concerning animals, as mentioned above it is only stating to be investigated in the allopolyploid context. In the animal genomes DNA methylation occurs preferentially at cytosines that are followed by guanines, called "CpGs. Methylation of the 5-position of cytosine (5mC) is mediated by DNA methyltransferases DNMT3a and DNMT3b and maintained in dividing cells by DNMT1 (Choleva-Waclaw *et al.,* 2016). For invertebrate genomes, methylation happens mostly at the gene bodies (exons and introns) while vertebrate genomes are heavily methylated, not only on the gene bodies but also and importantly in repetitive sequences as transposable elements. (Keller *et al.,* 2016).

1.6. Aims and structure of the thesis

The original findings of this work are enclosed in chapters 2 to 6, corresponding to 4 full articles already published in indexed international scientific journals and 1article submitted for publication. Due to the complexity of the models and to the overlap and interdependence of results, chapters succeed each other by the chronological order of publication, reflecting the progression of the work over time.

Over the last 20 years, the *S. alburnoides* hybrid complex has been used as a model system to study a variety of topics, from polyploidy (Gromicho *et al.,* 2006a, 2006b) to alternative reproductive strategies (Alves *et al.,* 1999) and sex determination and differentiation (Pala *et al.,* 2008a, 2009). It was also the first model used in the first studies tackling allopolyploid genome regulation and genome interactions in the vertebrate allopolyploid context (Pala et al., 2008b, 2010). Those previous studies in *S. alburnoides* complex (Pala *et al*., 2008a, 2010) have led to the assemble of an interesting theory of gene expression global dosage compensation by allele copy silencing operating in the allotriploids of this complex.

Yet, many questions remained unanswered in what concerns global gene expression regulation in *S. alburnoides* complex. It is still unknown whether dosage compensation acts throughout the whole genome, or if its occurrence is restricted,

specifically or randomly, to a subset of genes. Also, it remains to be fully demonstrated and understood if allelic silencing is happening globally in the transcriptome; if it is randomly copied or if there exists a genomic bias; or even if all three genome copies are expressed but with strikingly allelic imbalance.

So, the main goal of this thesis is to illustrate how a successful allopolyploid animal, the emblematic allopolyploid *Squalius alburnoides*, globally transcriptionally deals with the genomic stress derived from hybridization and polyploidy.

Although there was a parsimonious option of occurrence of differential expression regulation due to differential genome interactions (Pala *et al*., 2008a, 2010) to explain the irregular genome specific allelic silencing through the various forms and different tissues of the *S. alburnoides* complex, (Pala *et al.,* 2008a, 2010), other possibilities exist that would explain the observations. For example, the expression differences between individuals and/or between organs could be the result of mosaicism between organs and/or within an organ. Hence, before starting an expensive and labor-intensive pursuit for further clarifications, at chapter 2 this simple possibility has first been investigated and excluded as reason for the observations.

Another very significant void in the literature that this thesis aimed to clarify is whether the reported silencing mechanism in triploid *S. alburnoides, that is* very frequent among both natural and synthetized allopolyploid plants (Adams *et al*. 2003, 2004), is also a common mechanism among other natural and synthesized allopolyploid vertebrates. Obviously starting from other fish, artificially produced allopolyploid medakas (*Oryzias latipes*) and natural and artificially produced allopolyploid amazon mollies (*Poecilia formosa*) were used as models to address this question at chapter 3 and 5 respectively.

At chapter 4, RNA-seq Illumina sequencing was used to perform a first comparative transcriptomic analysis of *S. alburnoides* complex. Gene expression levels for diploid and triploid hybrids and of the parental genomic biotypes have been assessed and compared.

In this thesis, also the question of which mechanisms could be responsible for the reported allelic silencing described in the *S. alburnoides* system was addressed. In specific, at chapter 5, DNA methylation (5-mC) was evaluated as a possible candidate mechanism.

At chapter 6, allele specific quantification was performed on a genome wide scale and frequencies of complete allelic silencing and of unequal expression of alleles were identified in diploid and triploid *S. alburnoides*.

At chapter 7, the findings and partial discussions enclosed in each one of the previous 5 chapters (chapters 2 to 6) are compiled to provide an overview of the achievements of this work and the answers found to the several questions that were open at its beginning.

The last part of the present thesis (Chapter 8) corresponds to the enunciation of the main achievements that this work has put forward.

1.7. References

Abbott, R.**,** Albach, D., Ansell, S., Arntzen, J.W., Baird, S. J., … Zinner, D. (2013). Hybridization and speciation. *Journal of Evolutionary Biology*, 26, 229-246. https://doi.org/10.1111/j.1420-9101.2012.02599.x

Adams, K. L., Cronn, R., Percifield, R., and Wendel, J.F. (2003). Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(8), 4649–4654. http://doi.org/10.1073/pnas.0630618100

Adams, K.L., and Wendel, J.F. (2004). Exploring the genomic mysteries of polyploidy in cotton. *Biological Journal of the Linnean Society*, 82, 573-581. https://doi.org/10.1111/j.1095-8312.2004.00342.x

Adams, K.L., and Wendel, J.F. (2005). Polyploidy and genome evolution in plants. *Current Opinion Plant Biology*, 8, 135-141. https://doi.org/10.1016/j.pbi.2005.01.001

Alleman, M., and Doctor, J. (2000). Genomic imprinting in plants: observations and evolutionary implications. *Plant Molecular Biology*, 43,147-161. https://doi.org/10.1023/A:1006419025155

Alves, M.J., Coelho, M.M., Collares-Pereira, M.J., Dowling, T.E., (1997). Maternal ancestry of the Rutilus alburnoides complex (Teleostei, Cyprinidae) as determined by analysis of cytochrome b sequences. *Evolution*, 51, 1584 – 1592. https://doi.org/10.1111/j.1558-5646.1997.tb01481.x

Alves, M. J., Coelho, M.M., Próspero, M.I., and Collares-Pereira, M.J. (1999). Production of fertile unreduced sperm by hybrid males of the Rutilus alburnoides complex (Teleostei, cyprinidae). An alternative route to genome tetraploidization in unisexuals. *Genetics*, 151(1), 277–283.

Alves, M.J., Coelho, M.M., and Collares-Pereira, M.J. (2001). Evolution in action through hybridisation and polyploidy in an Iberian freshwater fish: a genetic review. *Genetica,* 111, 375-385. https://doi.org/10.1023/A:1013783029921

Alves, M.J., Coelho, M.M., Collares-Pereira, M.J., and Dowling, T.E. (1997). Maternal ancestry of the Rutilus alburnoides complex (Teleostei, Cyprinidae) as determined

by analysis of cytochrome b sequences. *Evolution*, 51, 1584-1592. https://doi.org/10.1111/j.1558-5646.1997.tb01481.x

Amores, A., Force, A., Yan, Y.L., Joly, L., Amemiya, C., … Postlethwait, J.H. (1998). Zebrafish hox clusters and vertebrate genome evolution. *Science*, 282, 1711-1714. http://science.sciencemag.org/content/282/5394/1711.long

Antonarakis, S.E., Lyle, R., Dermitzakis, E.T., Reymond, A., and Deutsch, S. (2004). Chromosome 21 and Down syndrome: From genomics to pathophysiology. *Nature Reviews Genetics*, 5, 725-738. https://doi.org/10.1038/nrg1448

Arai, R. (2011). Fish Karyotypes. A Check List. Springer, Tokyo, Japan. doi:10.1007/978-4-431-53877-6

Auger, D.L., Gray, A.D., Ream, T.S., Kato, A., Coe, E.H., and Birchler, J.A. (2005). Nonadditive gene expression in diploid and triploid hybrids of maize. *Genetics*, 169, 389-397. https://doi.org/10.1534/genetics.104.032987

Birchler, J. A., and Veitia, R. A. (2010). The Gene Balance Hypothesis: implications for gene regulation, quantitative traits and evolution. The New Phytologist, 186(1), 54–62. http://doi.org/10.1111/j.1469-8137.2009.03087.x

Birchler, J.A., and Veitia, R.A. (2007). The gene balance hypothesis: from classical genetics to modern genomics. *Plant Cell*, 19, 395-402. https://doi.org/10.1105/tpc.106.049338

Birchler, J.A., Yao, H., Chudalayandi, S., Vaiman, D., and Veitia, R.A. (2010). Heterosis. *Plant Cell*, 22, 2105-2112. https://doi.org/10.1105/tpc.110.076133

Blighe, K., DeDionisio, L., Christie, K. A., Chawes, B., Shareef, S., Kakouli-Duarte, T., … Moore, C. B. T. (2018). Gene editing in the context of an increasingly complex genome. *BMC Genomics*, 19, 595. http://doi.org/10.1186/s12864-018-4963-8

Boff, T., and Schifino-Wittmann, M.T. (2003). Segmental allopolyploidy and paleopolyploidy in species of Leucaena benth: evidence from meiotic behaviour analysis. *Hereditas*. 38(1), 27-35. https://doi.org/10.1034/j.1601-5223.2003.01646.x

Bogart, J.P., Bi, K., Fu, J.Z., Noble, D.W.A., and Niedzwiecki, J. (2007). Unisexual salamanders (genus Ambystoma) present a new reproductive mode for eukaryotes. *Genome,* 50, 119-136. https://doi.org/10.1139/G06-152

Bull, J.J. (1983). Evolution of sex determining mechanisms. Benjamin/Cummings Pub. Co., Advanced Book Program, Menlo Park, California, U.S.A. https://trove.nla.gov.au/work/21519895

Carmona, J.Á., Sanjur, O.I., Doadrio, I., Machordom, A., and Vrijenhoek, R.C. (1997). Hybridogenetic reproduction and maternal ancestry of polyploid Iberian fish: the Tropidophoxinellus alburnoides complex. *Genetics,* 146, 983-993. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1208066/

Chen, Z.J. (2007). Genetic and Epigenetic Mechanisms for Gene Expression and Phenotypic Variation in Plant Polyploids. *Annual review of plant biology*, 58, 377-406. Chen, Z. J. (2007). Genetic and Epigenetic Mechanisms for Gene Expression and Phenotypic Variation in Plant Polyploids. Annual Review of Plant Biology, 58, 377–406. http://doi.org/10.1146/annurev.arplant.58.032806.103835

Cholewa-Waclaw, J., Bird, A., von Schimmelmann, M., Schaefer, A., Yu, H., Song, H., … Tsai, L.-H. (2016). The Role of Epigenetic Mechanisms in the Regulation of Gene Expression in the Nervous System. *The Journal of Neuroscience*, 36(45), 11427–11434. http://doi.org/10.1523/JNEUROSCI.2492-16.2016

Clausen, J., Keck, D.D., and Hiesey, W.M. (1945). Plant evolution through amphiploidy and autoploidy, with examples from the Madiinae. Washington, DC, USA: Carnegie Institution of Washington.

Collares-Pereira, M.J., and Coelho, M.M. (2010). Reconfirming the hybrid origin and generic status of the Iberian cyprinid complex Squalius alburnoides. *Journal of Fish Biology*. 76, 707-715. https://doi.org/10.1111/j.1095-8649.2009.02460.x

Collares-Pereira, M.J., Matos, I., Morgado-Santos, M., and Coelho, M.M. (2013). Natural Pathways towards Polyploidy in Animals: The Squalius alburnoides Fish Complex as a Model System to Study Genome Size and Genome Reorganization in Polyploids. *Cytogenetic Genome Research*, 140, 97-116. https://doi.org/10.1159/000351729

Comai, L. (2005). The advantages and disadvantages of being polyploid. *Nature Reviews Genetics*, 6, 836-846. https://doi.org/10.1038/nrg1711

Covelo-Soto, L., Leunda, P.M., Pérez-Figueroa, A., and Morán, P. (2015). Genome-wide methylation study of diploid and triploid brown trout (Salmo trutta L.). *Animal Genetics*, 46, 280-288. https://doi.org/10.1111/age.12287

Crespo-López, M.E., Duarte, T., Dowling, T., and Coelho, M.M. (2006). Modes of reproduction of the hybridogenetic fish Squalius alburnoides in the Tejo and Guadiana rivers: an approach with microsatellites. *Zoology*, 109, 277-286. https://doi.org/10.1016/j.zool.2006.03.008

Cunha, C., Bastir, M., Coelho, M.M., and Doadrio, I. (2009). Body shape evolution among ploidy levels of the Squalius alburnoides hybrid complex (Teleostei, Cyprinidae). *Journal of Evolutionary Biology*, 22, 718-728. https://doi.org/10.1111/j.1420-9101.2009.01695.x

Cunha, C., Coelho, M.M., Carmona, and J.A., Doadrio, I. (2004). Phylogeographical insights into the origins of the Squalius alburnoides complex via multiple hybridization events. *Molecular Ecology*, 13, 2807-2817. https://doi.org/10.1111/j.1365-294X.2004.02283.x

Cunha, C., Doadrio, I., and Coelho, M.M. (2008). Speciation towards tetraploidization after intermediate processes of non-sexual reproduction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1505), 2921–2929. http://doi.org/10.1098/rstb.2008.0048

DePamphilis, M.L. (2016). Genome Duplication: The Heartbeat of Developing Organisms. *Current Topics in Developmental Biology*, 116, 201–229. http://doi.org/10.1016/bs.ctdb.2015.10.002

Dodsworth, S., Chase, M.W., and Leitch, A.R. (2016). Is post-polyploidization diploidization the key to the evolutionary success of angiosperms?. *Botanical Journal Linnean Society*, 180, 1-5. https://doi.org/10.1111/boj.12357

Doyle, J.J., and Sherman-Broyles, S. (2017). Double trouble: taxonomy and definitions of polyploidy. *New Phytologist*, 213, 487-493. https://doi.org/10.1111/nph.14276

Egozcue, S., Blano, J., Vidal, F., and Egozcue, J. (2002). Diploid sperm and the origin of triploidy. *Human Reproduction*, 17, 5-7. https://doi.org/10.1093/humrep/17.1.5

Evans, B.J., Pyron, R.A., Wiens, J.J. (2012). Polyploidization and sex chromosome evolution in amphibians. In: Soltis, P., Soltis, D. (eds) Polyploidy and Genome Evolution. Springer, Berlin, Heidelberg, Germany. https://doi.org/10.1007/978-3-642-31442-1_18

Gallardo, M.H., Bickham, J.W., Hoeycutt, R.L., Ojeda, R.A., and Köhler, N. (1999). Discovery of tetraploidy in a mammal. *Nature* ,401, 341. https://doi.org/10.1038/43815

Gallardo, M.H., González, C.A., and Cebrián, I. (2006). Molecular cytogenetics and allotetraploidy in the red vizcacha rat, Tympanoctomys barrerae (Rodentia, Octodontidae). *Genomics*, 88, 214-221. https://doi.org/10.1016/j.ygeno.2006.02.010

Gallardo, M.H., Kausel, G., Jiménez, A., Bacquet, C., González, C., Figueroa, J., Köhler, N., and Ojeda, R. (2004). Whole-genome duplications in South American desert rodents (Octodontidae). *Biological Journal of the Linnean Society*, 82, 443-451. https://doi.org/10.1111/j.1095-8312.2004.00331.x

Grandont, L., Jenczewski, E., and Lloyd, A. (2013). Meiosis and its deviations in polyploid plants. *Cytogenetic and Genome Research*, 140, 171-84. https://doi.org/10.1159/000351730

Greaves, I. K., Gonzalez-Bayon, R., Wang, L., Zhu, A., Liu, P.-C., Groszmann, M., … Dennis, E. S. (2015). Epigenetic Changes in Hybrids. Plant Physiology, 168(4), 1197–1205. http://doi.org/10.1104/pp.15.00231

Gregory, T.R., and Mable, B.K. (2005). Polyploidy in animals. The Evolution of the Genome, Elsevier Incorporation, Burlington. https://doi.org/10.1016/B978-012301463-4/50010-3

Grishanin, A.K., Rasch, E.M., Dodson, S.I., and Wyngaard, G.A. (2006). Genetic architecture of the cryptic species complex of Acanthocyclops vernalis (Crustacea: Copepoda). II. Crossbreeding experiments, cytogenetics, and a model of chromosomal evolution. *Evolution*, 60, 247-256. https://www.jstor.org/stable/4095213

Gromicho, M., and Collares-Pereira, M.J. (2007). The evolutionary role of hybridization and polyploidy in an Iberian cyprinid fish – a cytogenetic review, in

Pisano, E., Ozouf-Costaz, Cm., Foresti, F., Kapoor, B.G. Fish Cytogenetics, Science Publishers, Enfield, U.S.A.

Gromicho, M., Coelho, M.M., Alves, M.J., and Collares-Pereira, M.J. (2006a). Cytogenetic analysis of Anaecypris hispanica and its relationship with the paternal ancestor of the diploid-polyploid Squalius alburnoides complex. *Genome,* 49(12), 1621-7. https://doi.org/10.1139/g06-121

Gromicho, M., Coutanceau, J.P., Ozouf-Costaz, C., and Collares-Pereira, M.J. (2006). Contrast between extensive variation of 28S rDNA and stability of 5S rDNA and telomeric repeats in the diploid-polyploid Squalius alburnoides complex and in its maternal ancestor Squalius pyrenaicus (Teleostei, Cyprinidae). *Chromosome Research*, 14(3), 297-306. https://doi.org/10.1007/s10577-006-1047-4

Ha, M., Lu, J., Tian, L., Ramachandran, V., Kasschau, K. D., Chapman, E. J., … Chen, Z. J. (2009). Small RNAs serve as a genetic buffer against genomic shock in Arabidopsis interspecific hybrids and allopolyploids. *Proceedings of the National Academy of Sciences of the United States of America*, 106(42), 17835–17840. http://doi.org/10.1073/pnas.0907003106

Hall, W.P. (2009). Chromosome variation, genomics, speciation and evolution in Sceloporus lizards. *Cytogenetic and Genome Research,* 127, 143-165 (2009). https://doi.org/10.1159/000304050

Harrison, R.G., and Larson, E.L. (2014). Hybridization, Introgression, and the Nature of Species Boundaries. *Journal of Heredity*, 105, 795–809. https://doi.org/10.1093/jhered/esu033

Hegarty, M., Coate, J., Sherman-Broyles, S., Abbott, R., Hiscock, S., and Doyle, J. (2013). Lessons from natural and artificial polyploids in higher plants. *Cytogenetic and Genome Research*, 140, 204-25. https://doi.org/10.1159/000353361

Hegarty, M.J., Barker, G.L., Wilson, I.D., Abbott, R.J., Edwards, K.J., and Hiscock, S.J. (2006). Transcriptome shock after interspecific hybridization in Senecio is ameliorated by genome duplication. *Current Biology*, 16, 1652-1659. https://doi.org/10.1016/j.cub.2006.06.071

Huntzinger, E., and Izaurralde, E. (2011). Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nature Reviews Genetics*, 12, 99-110. https://doi.org/10.1038/nrg2936

Hyodo-Taguchi, Y. (1980). Establishment of inbred strains of the teleost Oryzias latipes. *Zoological magazine (Tokyo)*, 89, 283–301.

Inácio, A., Pinho, J., Pereira, P.M., Comai, L., and Coelho, M.M. (2012). Global Analysis of the Small RNA Transcriptome in Different Ploidies and Genomic Combinations of a Vertebrate Complex – The Squalius alburnoides. *PLoS ONE*, 7(7), e41158. http://doi.org/10.1371/journal.pone.0041158

Ishikawa, Y., M. Yoshimoto, N., Yamamoto, and H. Ito. (1999). Different brain morphologies from different genotypes in a single teleost species, the medaka (Oryzias latipes). Brain Behavior and Evolution, 53, 2–9. https://doi.org/10.1159/000006577

Jackson, R.C., and Jackson, J.W. (1996). Gene segregation in autotetraploids: prediction from meiotic configurations. *American Journal of Botany,* 83, 673–678. https://www.jstor.org/stable/2445844

Jackson, S. A. (2017). Epigenomics: dissecting hybridization and polyploidization. *Genome Biology*, 18, 117. http://doi.org/10.1186/s13059-017-1254-7

Jenczewski, E., and Alix, K. (2004). From diploids to allopolyploids: the emergence of efficient pairing control genes in plants. *Critical Reviews in Plant Sciences*, 23: 21–45. https://doi.org/10.1080/07352680490273239

Jiang, Q., Li, Q., Yu, H., and Kong, L.F. (2016). Inheritance and variation of genomic DNA methylation in diploid and triploid Pacific Oyster (Crassostrea gigas). *Marine Biotechnology*, 18, 124-132. https://link.springer.com/article/10.1007%2Fs10126-015-9674-4

Johnson, R.M., Shrimpton, J.M., Cho, G.K., and Heath, D.D. (2007). Dosage effects on heritability and maternal effects in diploid and triploid Chinook salmon (Oncorhynchus tshawytscha). *Heredity*, 98, 303-310
https://www.nature.com/articles/6800941

Kasahara, M., Naruse, K., Sasaki, S., Nakatani, Y., Qu, W., Ahsan, B., … Kohara, Y. (2007). The medaka draft genome and insights into vertebrate genome evolution. *Nature,* 447, 714–719 (2007). https://www.nature.com/articles/nature05846

Keller, T. E., Han, P., and Yi, S. V. (2016). Evolutionary Transition of Promoter and Gene Body DNA Methylation across Invertebrate–Vertebrate Boundary. *Molecular Biology and Evolution*, 33(4), 1019–1028. http://doi.org/10.1093/molbev/msv345

Kimura, T., Shimada, A., Sakai, N., Mitani, H., Naruse, K., Takeda, H., … Shinya, M. (2007). Genetic Analysis of Craniofacial Traits in the Medaka. *Genetics*, 177(4), 2379–2388. http://doi.org/10.1534/genetics.106.068460

Kirchmaier, S., Naruse, K., Wittbrodt, J., and Loosli, F. (2015). The Genomic and Genetic Toolbox of the Teleost Medaka (Oryzias latipes). *Genetics,* 199(4), 905–918. http://doi.org/10.1534/genetics.114.173849

Koroma, A.P., Jones, R., and Michalak, P. (2011). Snapshot of DNA methylation changes associated with hybridization in Xenopus. *Physiological Genomics*, 43(22), 1276-80. https://doi.org/10.1152/physiolgenomics.00110.2011

Lamatsch, D.K., Stöck, M., Fuchs, R., Döbler, M., Wacker, R., Parzefall, J., Schlupp, I. and Schartl, M. (2010). Morphology, testes development and behaviour of unusual triploid males in microchromosome-carrying clones of Poecilia formosa. *Journal of Fish Biology*, 77, 1459-1487. https://doi.org/10.1111/j.1095-8649.2010.02766.x

Lampert, K., and Schartl, M. (2008). The origin and evolution of a unisexual hybrid: Poecilia formosa. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1505), 2901–2909. http://doi.org/10.1098/rstb.2008.0040

Lampert, K.P., Lamatsch, D.K., Epplen, J.T., and Schartl, M. (2005). Evidence for a monophyletic origin of triploid clones of the Amazon molly, Poecilia formosa. *Evolution*, 59(4), 881-9. https://doi.org/10.1554/04-453

Lampert, K.P., Lamatsch, D.K., Fischer, P., Epplen, J.T., Nanda, I., Schmid, M., and Schartl, M. (2007). Automictic reproduction in interspecific hybrids of poeciliid fish. *Current Biology*, 17(22), 1948-1953. https://doi.org/10.1016/j.cub.2007.09.064

Landergott, U., Nacir,i Y., Schneller, J.J., and Holderegger, R. (2006). Allelic configuration and polysomic inheritance of highly variable microsatellites in

tetraploid gynodioecious Thymus praecox agg. *Theoretical and Applied Genetics*, 113, 453–465. https://doi.org/10.1007/s00122-006-0310-6

Le Comber, S.C., and Smith, C. (2004). Polyploidy in fishes: patterns and processes. *Biological Journal of the Linnean Society*, 82, 431-442. https://doi.org/10.1111/j.1095-8312.2004.00330.x

Leggatt, R.A., and Iwama, G.K. (2003). Occurrence of polyploidy in fishes. *Reviews in Fish Biology and Fisheries*, 13, 237-246. https://doi.org/10.1023/B:RFBF.0000033049.00668.fe

Li, Z., Lu, X., Gao, Y., Liu, S., Tao, M., Xiao, H., … Luo, J. (2011). Polyploidization and epigenetics. *Chinese Science Bulletin*, 56, 245−252. https://doi.org/10.1007/s11434-010-4290-1

Lim, L.P., Lau, N.C., Garrett-Engele, P., Grimson, A., Schelter, J.M., Castle, J., Bartel, D.P., Linsley, P.S., and Johnson J.M. (2005). Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, 433, 769-773. https://www.nature.com/articles/nature03315

Lowe, A.J., and Abbott, R.J. (2015). Hybrid swarms: catalysts for multiple evolutionary events in Senecio in the British Isles. *Plant Ecology & Diversity*, 8, 449-463. https://doi.org/10.1080/17550874.2015.1028113

Mable, B. K. (2013). Polyploids and hybrids in changing environments: winners or losers in the struggle for adaptation? *Heredity*, 110(2), 95–96. http://doi.org/10.1038/hdy.2012.105

Mable, B.K. (2004). 'Why polyploidy is rarer in animals than in plants': myths and mechanisms. *Biological Journal of the Linnean Society*, 82, 453-466. https://doi.org/10.1111/j.1095-8312.2004.00332.x

Mable, B.K., Alexandrou, M.A. and Taylor, M.I. (2011). Genome duplication in amphibians and fish: an extended synthesis. *Journal of Zoology*, 284, 151-182. https://doi.org/10.1111/j.1469-7998.2011.00829.x

Machado, M.P., Matos, I., Grosso, A.R., Schartl, M. and Coelho, M.M. (2016). Non-canonical expression patterns and evolutionary rates of sex-biased genes in a seasonal fish. *Molecular Reproduction and Development*, 83, 1102-1115. doi:10.1002/mrd.22752

Mallet, J. (2007). Hybrid speciation. *Nature* 446, 279-283. https://www.nature.com/articles/nature05706

Malone, J.H., Cho, D.-Y., Mattiuzzo, N.R., Artieri, C.G., Jiang, L., Dale, R.K., … Oliver, B. (2012). Mediation of Drosophila autosomal dosage effects and compensation by network interactions. *Genome Biology*, 13(4), R28. http://doi.org/10.1186/gb-2012-13-4-r28

Mank, J.E., and Avise, J.C. (2009). Evolutionary Diversity and Turn-Over of Sex Determination in Teleost Fishes. *Sexual Development*, 3, 60-67. https://doi.org/10.1159/000223071

Marques, I., Loureiro, J., Draper, D., Castro, M., Castro, S., and Arroyo, J. (2018), How much do we know about the frequency of hybridisation and polyploidy in the Mediterranean region?. *Plant Biology Journal*, 20, 21-37. https://doi.org/10.1111/plb.12639

Matos, I., Machado, M.P., Sucena, E., Collares-Pereira, M.J., Schartl, M., and Coelho, M.M. (2010). Evidence for hermaphroditism in the Squalius alburnoides allopolyploid fish complex. *Sexual Development*, 4,170-175. https://doi.org/10.1159/000313359

Matsuda, M., and Sakaizumi, M. (2015). Evolution of the sex-determining gene in the teleostean 495 genus Oryzias. *General and Comparative Endocrinology*, 239, 80-88. https://doi.org/10.1016/j.ygcen.2015.10.004

Matsuda, M., Yonekawa, H., Hamaguchi, S., Sakaizumi, M. (1997). Geographic variation and diversity in the mitochondrial DNA of the medaka, Oryzias latipes, as determined by restriction endonuclease analysis. *Zoological Science*, 14, 517–526. https://doi.org/10.2108/zsj.14.517

McManus, C.J., Coolon, J.D., Duff, M.O., Eipper-Mains, J., Graveley, B.R., and Wittkopp, P.J. (2010). Regulatory divergence in Drosophila revealed by mRNA-seq. *Genome Research*, 20(6), 816–825. http://doi.org/10.1101/gr.102491.109

Meyer, A., and Schartl, M. (1999). Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. Current Opinion in Cell Biology, 11(6), 699-704. https://doi.org/10.1016/S0955-0674(99)00039-3

Morgado-Santos, M., Carona, S., Magalhães, M.F., Vicente, L., and Collares-Pereira, M.J. (2016). Reproductive dynamics shapes genomotype composition in an allopolyploid complex. *Proceedings of the Royal Society B: Biological Sciences*, 283(1831), 20153009. http://doi.org/10.1098/rspb.2015.3009

Morgado-Santos, M., Carona, S., Vicente, L., and Collares-Pereira, M. J. (2017). First empirical evidence of naturally occurring androgenesis in vertebrates. *Royal Society Open Science*, 4(5), 170200. http://doi.org/10.1098/rsos.170200

Müller, H.J. (1925). Why polyploidy is rarer in animals than in plants. *The American Naturalist*, 59, 346-353.

Nanda, I, Schartl, M., Feichtinger, W., Schlupp, I., Parzefall, J. and Schmid, M. (1995). Chromosomal evidence for laboratory synthesis of a triploid hybrid between the gynogenetic teleost Poecilia formosa and its host species. *Journal of Fish Biology*, 47, 619-623. https://doi.org/10.1111/j.1095-8649.1995.tb01928.x

Nanda, I., Schlupp, I., Lamatsch, D.K., Lampert, K.P., Schmid, M., and Schartl, M. (2007). Stable Inheritance of Host Species-Derived Microchromosomes in the Gynogenetic Fish Poecilia formosa. *Genetics*, 177(2), 917–926. http://doi.org/10.1534/genetics.107.076893

Naruse, K., Naruse, K., Tanaka, M., Mita, K., Shima, A., Postlethwait, J., and Mitani, H. (2004). A Medaka Gene Map: The Trace of Ancestral Vertebrate Proto-Chromosomes Revealed by Comparative Gene Mapping. *Genome Research*, 14(5), 820–828. http://doi.org/10.1101/gr.2004004

Naruse, K., Tanaka, M., and Takeda, H. (2011). Medaka: A Model for Organogenesis, Human Disease, and Evolution, Springer, Tokyo, Japan. doi10.1007/978-4-431-92691-7

Nieto Feliner, G., Álvarez, I., Fuertes-Aguilar, J., Heuertz, M., Marques, I., Moharrek, F., … Villa-Machío, I. (2017). Is homoploid hybrid speciation that rare? An empiricist's view. *Heredity*, 118(6), 513–516. http://doi.org/10.1038/hdy.2017.7

Ohno, S. (1970). Evolution by gene duplication. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-86659-3

Orr, A.H. (1990). "Why Polyploidy is Rarer in Animals Than in Plants" Revisited. *The American Naturalist*, 136, 6, 759-770. https://doi.org/10.1086/285130

Orr, D.W. (1990). *Conservation Biology*, 4, 219-220. https://doi.org/10.1111/j.1523-1739.1990.tb00279.x

Otto, S.P. (2007). The evolutionary consequences of polyploidy. *Cell* ,131, 452-462. https://doi.org/10.1016/j.cell.2007.10.022

Otto, S.P., Whitton, J. (2000). Polyploid The Virtue of Conservation Education incidence and evolution. *Annual Review of Genetics*, 34, 401-437. https://doi.org/10.1146/annurev.genet.34.1.401

Over, R. S., and Michaels, S. D. (2014). Open and Closed: The Roles of Linker Histones in Plants and Animals. *Molecular Plant*, 7(3), 481–491. http://doi.org/10.1093/mp/sst164

Pala, I., and Coelho, M.M. (2005). Contrasting views over a hybrid complex: Between speciation and evolutionary 'dead-end'. *Gene*, 347, 283-294. https://doi.org/10.1016/j.gene.2004.12.010

Pala, I., Coelho, M.M., and Schartl, M. (2008a). Dosage compensation by gene-copy silencing in a triploid hybrid fish. Current Biology, 18, 1344-1348. https://doi.org/10.1016/j.cub.2008.07.096

Pala, I., Klüver, N., Thorsteinsdóttir, S., Schartl, M., and Coelho, M.M. (2008b). Expression pattern of antiMüllerian hormone (amh) in the hybrid fish complex of Squalius alburnoides. *Gene*, 410: 249–258 https://doi.org/10.1016/j.gene.2007.12.018

Pala, I., Schartl, M., Brito, M., Vacas, J.M., and Coelho, M.M. (2010). Gene expression regulation and lineage evolution: the North and South tale of the hybrid polyploid Squalius alburnoides complex. *Proceedings of the Royal Society B: Biological Sciences,* 277(1699), 3519–3525. http://doi.org/10.1098/rspb.2010.1071

Parisod, C., Alix, K., Just, J., Petit, M., Sarilar, V., Mhiri, C., Ainouche, M., Chalhoub, B., and Grandbastien, M. (2010). Impact of transposable elements on the organization and function of allopolyploid genomes. *New Phytologist*, 186, 37-45. https://doi.org/10.1111/j.1469-8137.2009.03096.x

Pignatta, D., and Comai, L. (2009). Parental squabbles and genome expression: lessons from the polyploids. *Journal of Biology*, 8(4), 43. http://doi.org/10.1186/jbiol140

Ráb, P., and Collares-Pereira, M.J. (1995). Chromosomes of European cyprinid fishes (Cyprinidae, Cypriniformes): a review. *Folia Zoologica*, 44, 193-214.

Ramsey, J., and Schemske, D.W. (1998). Pathways, Mechanisms, and Rates of Polyploidy Formation in Flowering Plants. *Annual Review of Ecology and Systematics*, 29, 467-501. https://doi.org/10.1146/annurev.ecolsys.29.1.467

Ramsey, J., and Schemske, D.W. (2002). Neopolyploidy in flowering plants. *Annual Review of Ecology and Systematics*, 33, 589–639.
https://doi.org/10.1146/annurev.ecolsys.33.010802.150437

Rapp, R. A., Udall, J.A., and Wendel, J.F. (2009). Genomic expression dominance in allopolyploids. *BMC Biology*, 7, 18. http://doi.org/10.1186/1741-7007-7-18

Renner, S., and Ricklefs, R. Dioecy and Its Correlates in the Flowering Plants. (1995). *American Journal of Botany*, 82(5), 596-606 https://www.jstor.org/stable/2445418

Ribeiro, F., Cowx, I.G., Tiago, P., Filipe, A.F., Moreira da Costa L, and Collares-Pereira, M.J. (2003). Growth and reproductive traits of diploid and triploid forms of Squalius alburnoides cyprinid complex in a tributary of Guadiana River, Portugal. *Archives für Hydrobiology*, 156, 471-484. doi: 10.1127/0003-9136/2003/0156-0471

Robalo, J.I., Sousa Santos, C., Levy, A., Almada, V.C. (2006). Molecular insights on the taxonomic position of the paternal ancestor of the Squalius alburnoides hybridogenetic complex. *Molecular Phylogenetics and Evolution*, 39, 276-281.

Sakaizumi, M., Moriwaki, K., and Egami, N. (1983). Allozymic variation and regional differentiation in wild population of the fish Oryzias latipes. *Copeia*, 311–318 (1983). https://www.jstor.org/stable/1444373

Sattler, M.C., Carvalho, C.R. and Clarindo, W.R. (2016). *Planta,* 243, 281. https://doi.org/10.1007/s00425-015-2450-x

Schartl, A., Hornung, U., Nanda, I., Wacker, R., Müller-Hermelink, H. K., Schlupp, I., … Schartl, M. (1997). Susceptibility to the development of pigment cell tumors in a clone of the Amazon molly, Poecilia formosa, introduced through a microchromosome. *Cancer Research*. 57, 2993-3000.
http://cancerres.aacrjournals.org/content/57/14/2993.long

Schartl, M. (2014). Beyond the zebrafish: diverse fish species for modeling human disease. *Disease Models & Mechanisms*, 7(2), 181–192. http://doi.org/10.1242/dmm.012245

Schlupp, I. (2010). Mate choice and the Amazon molly: how sexuality and unisexuality can coexist. *Journal of Heredity*, 101, S55-61 https://doi.org/10.1093/jhered/esq015

Schories, S., Lampert, K.P., Lamatsch, D.K., de León, F.J.G., and Schartl, M. (2007). Analysis of a possible independent origin of triploid P. formosa outside of the Río Purificación river system. *Frontiers in Zoology*, 4, 13. http://doi.org/10.1186/1742-9994-4-13

Shima, A., and Mitani, H. (2004). Medaka as a research organism: past, present and future. (2004). *Mechanisms of Development*, 121(7-8), 599-604. https://doi.org/10.1016/j.mod.2004.03.011

Soltis, P.S., and Soltis, D.E. (2012). Polyploidy and Genome Evolution. Springer-Verlag Berlin Heidelberg. Doi:10.1007/978-3-642-31442-1

Sousa-Santos, C., Collares-Pereira, M.J., and Almada, V. (2007a). Reading the history of a hybrid fish complex from its molecular record. *Molecular Phylogenetics and Evolution*, 45, 981-996. https://doi.org/10.1016/j.ympev.2007.05.011

Sousa-Santos, C., Collares-Pereira, M.J., and Almada, V. (2007b). Fertile triploid males—an uncommon case among hybrid vertebrates. *Journal of Experimental Zoology*, 307A, 220-225. https://doi.org/10.1002/jez.363

Sousa-Santos, C., Collares-Pereira, M.J., and Almada, V.C. (2006a). Evidence of extensive mitochondrial introgression with nearly complete substitution of the typical Squalius pyrenaicus-like mtDNA of the Squalius alburnoides complex (Cyprinidae) in an independent Iberian drainage. *Journal of Fish Biology*, 68, S292-S301. https://doi.org/10.1111/j.0022-1112.2006.01081.x

Sousa-Santos, C., Collares-Pereira, M.J., and Almada, V.C. (2006b). May a hybridogenetic complex regenerate the nuclear genome of both sexes of a missing ancestor? - First evidence on the occurrence of a nuclear non-hybrid Squalius alburnoides (Cyprinidae) female based on DNA sequencing. Journal of Natural History, 40, 1443-1448. doi: 10.1080/00222930600934111

Spivakov, M., Auer, T.O., Peravali, R., Dunham, I., Dolle, D., Fujiyama, A., … Wittbrodt, J. (2014). Genomic and Phenotypic Characterization of a Wild Medaka Population: Towards the Establishment of an Isogenic Population Genetic Resource in Fish. *G3: Genes|Genomes|Genetics*, 4(3), 433–445. http://doi.org/10.1534/g3.113.008722

Stebbins, G.L. (1950). Variation and Evolution in Plants. Columbia University Press, New York, USA.

Stebbins, G.L. (1971). Chromosomal Evolution in Higher Plants. Edward Arnold Ldt., London, UK.

Stebbins, G.L., Matzke, E.B., and Epling, C. (1947), Hybridization in a population of Quercus marilandica and Quercus ilicifolia. *Evolution*, 1, 79-88. https://doi.org/10.1111/j.1558-5646.1947.tb02716.x

Stenberg, P., and Saura, A. (2013). Meiosis and its deviations in polyploid animals. *Cytogenetic and Genome Research*, 140, 185-203. https://doi.org/10.1159/000351731

Stift, M., Berenos, C., Kuperus, P., and van Tienderen, P.H. (2008). Segregation Models for Disomic, Tetrasomic and Intermediate Inheritance in Tetraploids: A General Procedure Applied to Rorippa (Yellow Cress) Microsatellite Data. *Genetics*, 179(4), 2113–2123. http://doi.org/10.1534/genetics.107.085027

Stöck, M., and Lamatsch, D.K. (2013). Why comparing polyploidy research in animals and plants? *Cytogenetic and Genome Research*, 140(2-4), 75-8. https://doi.org/10.1159/000353304.

Svartman, M., Stone, G., and Stanyon, R. (2005). Molecular cytogenetics discards polyploidy in mammals. *Genomics* 85:425-430 https://doi.org/10.1016/j.ygeno.2004.12.004.

Takeda, H., and Shimada, A. (2010). The art of medaka genetics and genomics: what makes them so unique?. *Annual Reviews Genetics*, 44, 217-41. https://www.annualreviews.org/doi/10.1146/annurev-genet-051710-151001

Takehana, Y., Nagai, N., Matsuda, M., Tsuchiya, K., and Sakaizumi, M. (2003). Geographic variation and diversity of the cytochrome b gene in Japanese wild

populations of medaka, Oryzias latipes. *Zoological Sciences*, 20, 1279–1291. https://doi.org/10.2108/zsj.21.483

Takehana, Y., Naruse, K., and Sakaizumi, M. (2005). Molecular phylogeny of the medaka fishes genus Oryzias (Beloniformes: Adrianichthyidae) based on nuclear and mitochondrial DNA sequences. *Molecular Phylogenetics and Evolution*, 36, 417–428 https://doi.org/10.1016/j.ympev.2005.01.016

Takehana, Y., Uchiyama, S., Matsuda, M., Jeon, S.R., and Sakaizumi, M. (2004). Geographic variation and diversity of the cytochrome b gene in wild populations of medaka (Oryzias latipes) from Korea and China. *Zoological Science*, 21, 483–491. https://doi.org/10.2108/zsj.21.483

Tobler, M., and Schlupp, I. (2005). Parasites in sexual and asexual mollies (Poecilia, Poeciliidae, Teleostei): a case for the Red Queen? *Biology Letters*, 1(2), 166–168. http://doi.org/10.1098/rsbl.2005.0305

Tsigenopoulos, C.S., Ráb, P., Naran, D., and Berrebi, P. (2002). Multiple origins of polyploidy in the phylogeny of southern African barbs (Cyprinidae) as inferred from mtDNA markers. *Heredity*, 88(6), 466-73 https://doi.org/10.1038/sj.hdy.6800080

Uwa, H., and Ojima, Y. (1981). Detailed and banding karyotype analyses of the medaka, Oryzias latipes in cultured cells. *Proceedings of the Japan Academy, Series B*, 57, 39–43. https://doi.org/10.2183/pjab.57.39

Vogel, G. (1998). Doubled genes may explain fish diversity. *Science*, 281, 1119–1121. http://science.sciencemag.org/content/281/5380/1119

Wang, J., Tian, L., Lee, H.-S., Wei, N. E., Jiang, H., Watson, B., … Chen, Z. J. (2006). Genomewide Nonadditive Gene Regulation in Arabidopsis Allotetraploids. *Genetics*, 172(1), 507–517. http://doi.org/10.1534/genetics.105.047894

Warren, W.C., García-Pérez, R., Xu, S., Lampert, K.P., Chalopin, D., Stöck, M., … Schartl, M. (2018). Clonal polymorphism and high heterozygosity in the celibate genome of the Amazon molly. *Nature Ecology & Evolution*, 2, 669–679 https://doi.org/10.1038/s41559-018-0473-y

Wakamatsu, Y. (2008). Novel method for the nuclear transfer of adult somatic cells in medaka fish (Oryzias latipes): Use of diploidized eggs as recipients. *Development, Growth & Differentiation*, 50, 427-436.

https://doi.org/10.1111/j.1440-169X.2008.01050.x

Wertheim, B., Beukeboom, L.W., and van de Zande, L. (2013). Polyploidy in animals: effects of gene expression on sex determination, evolution and ecology. *Cytogenetic and Genome Research*, 140(2-4), 256-69. https://doi.org/10.1159/000351998

White, M.J.D. (1978). Modes of Speciation. W.H. Freeman, San Francisco, U.S.A.

Wittbrodt, J., Shima, A., and Schartl, M. (2002). Medaka--a model organism from the far East. *Nature Reviews Genetics*. 3(1):53-64. https://doi.org/10.1038/nrg704

Woodhead, A.D., Setlow, R.B., and Pond, V. (1984). The Amazon molly, Poecilia formosa, as a test animal in carcinogenicity studies: chronic exposures to physical agents. *National Cancer Institute monograph, 65*, 45-52.

Wright, K. M., Arnold, B., Xue, K., Šurinová, M., O'Connell, J., and Bomblies, K. (2015). Selection on Meiosis Genes in Diploid and Tetraploid Arabidopsis arenosa. *Molecular Biology and Evolution*, 32(4), 944–955. http://doi.org/10.1093/molbev/msu398

Xiao, J., Song, C., Liu, S., Tao, M., Hu, J., Wang, J., … Liu, Y. (2013). DNA Methylation Analysis of Allotetraploid Hybrids of Red Crucian Carp (Carassius auratus red var.) and Common Carp (Cyprinus carpio L.). PLoS ONE, 8(2), e56409. http://doi.org/10.1371/journal.pone.0056409

Yadav, T., Quivy, J.P., and Almouzni, G. (2018). Chromatin plasticity: A versatile landscape that underlies cell fate and identity. *Science*, 361(6409),1332-1336. doi: 10.1126/science.aat8950

Yakimowski, S.B. and Rieseberg, L.H. (2014). The role of homoploid hybridization in evolution: A century of studies synthesizing genetics and ecology. *American Journal of Botany*, 101: 1247-1258. https://doi.org/10.3732/ajb.1400201

Zhou, L., and Gui, J. (2017). Natural and artificial polyploids in aquaculture. *Aquaculture and Fisheries*, 2, 103-111. https://doi.org/10.1016/j.aaf.2017.04.003

Zhu, B., and Ge, W. (2018). Genome editing in fishes and their applications. General and Comparative Endocrinology. 257, 3-12. https://doi.org/10.1016/j.ygcen.2017.09.011

CHAPTER 2

Ploidy mosaicism and allele-specific gene expression differences in the allopolyploid *Squalius alburnoides*

Matos I, Sucena É, Machado MP, Gardner R, Inácio Â, Schartl M, Coelho MM. Ploidy mosaicism and allele-specific gene expression differences in the allopolyploid Squalius alburnoides.BMC Genetics, 12-101 (2011)

Matos et al. BMC Genetics 2011, 12:101
http://www.biomedcentral.com/1471-2156/12/101

BMC Genetics

RESEARCH ARTICLE Open Access

Ploidy mosaicism and allele-specific gene expression differences in the allopolyploid *Squalius alburnoides*

Isa Matos[1,2*], Élio Sucena[3,4], Miguel P Machado[1], Rui Gardner[3], Ângela Inácio[1], Manfred Schartl[2] and Maria M Coelho[1]

Abstract

Background: *Squalius alburnoides* is an Iberian cyprinid fish resulting from an interspecific hybridisation between *Squalius pyrenaicus* females (P genome) and males of an unknown *Anaecypris hispanica*-like species (A genome). *S. alburnoides* is an allopolyploid hybridogenetic complex, which makes it a likely candidate for ploidy mosaicism occurrence, and is also an interesting model to address questions about gene expression regulation and genomic interactions. Indeed, it was previously suggested that in *S. alburnoides* triploids (PAA composition) silencing of one of the three alleles (mainly of the P allele) occurs. However, not a whole haplome is inactivated but a more or less random inactivation of alleles varying between individuals and even between organs of the same fish was seen. In this work we intended to correlate expression differences between individuals and/or between organs to the occurrence of mosaicism, evaluating if mosaics could explain previous observations and its impact on the assessment of gene expression patterns.

Results: To achieve our goal, we developed flow cytometry and cell sorting protocols for this system generating more homogenous cellular and transcriptional samples. With this set-up we detected 10% ploidy mosaicism within the *S. alburnoides* complex, and determined the allelic expression profiles of ubiquitously expressed genes (*rpl8*; *gapdh* and *β-actin*) in cells from liver and kidney of mosaic and non-mosaic individuals coming from different rivers over a wide geographic range.

Conclusions: Ploidy mosaicism occurs sporadically within the *S. alburnoides* complex, but in a frequency significantly higher than reported for other organisms. Moreover, we could exclude the influence of this phenomenon on the detection of variable allelic expression profiles of ubiquitously expressed genes (*rpl8*; *gapdh* and *β-actin*) in cells from liver and kidney of triploid individuals. Finally, we determined that the expression patterns previously detected only in a narrow geographic range is not a local restricted phenomenon but is pervasive in rivers where *S. pyrenaicus* is sympatric with *S. alburnoides*.
We discuss mechanisms that could lead to the formation of mosaic *S. alburnoides* and hypothesise about a relaxation of the mechanisms that impose a tight control over mitosis and ploidy control in mixoploids.

Background

The chromosome theory of heredity rests on the consistency and stability of chromosome number and composition [1]. This consistency and stability is achieved by the existence of extremely precise and tightly controlled mechanisms of chromosome replication and segregation during cell divisions [2]. However, genetic information and the way it is inherited are not so invariant and rigorous as previously thought [3]. Experimental findings in reproductive genetics have shown that basic processes such as mitosis, meiosis/gametogenesis, fertilization and embryogenesis are often imprecise and present some level of plasticity [4]. It is through this mechanistic plasticity and the ability of organisms to cope with seemingly low frequencies of genetic aberrations that

* Correspondence: immatos@fc.ul.pt
[1]Centro de Biologia Ambiental, Departamento de Biologia Animal, Faculdade de Ciências da Universidade de Lisboa, Campo Grande, Lisbon, 1749-016, Portugal
Full list of author information is available at the end of the article

hybridization and polyploidy emerge as naturally occurring phenomena. In this light, allopolyploids, like the cyprinid fish *Squalius alburnoides*, constitute a paradigmatic example of successful escapers from the canonical rules of reproductive biology and heredity [5-9].

The *Squalius alburnoides* complex is endemic from the Iberian Peninsula. It resulted from interspecific hybridisation between females of *Squalius pyrenaicus* (P genome) and males of an unknown species related to *Anaecypris hispanica* (A genome) [reviewed in [10]].

S. alburnoides is described as an allopolyploid hybridogenetic complex, where allopolyploid refers to an increased ploidy level and hybrid genome composition of particular forms within the system; hybridogenetic refers to an alternative mode of reproduction; and complex is the technical terminus denoting a natural system composed of parental species and their hybrids, with altered modes of reproduction and reproductive interdependence [10].

Presently, and due to the altered reproductive modes adopted by *S. alburnoides* and the reproductive relationship established with several allopatric bisexual *Squalius* species, mainly *S. carolitertii* (C genome) and *S. pyrenaicus*, a multitude of ploidy levels and genomic constitutions can be found [10]. These include diploids (PA, CA), triploids (PAA, PPA, CAA, CCA) and tetraploids (PPAA, CCAA) depending on the geographical location (Additional file 1, Figure S1). In the Iberian southern basins an additional form is present, composed exclusively of males designated as "nuclear non-hybrid AA's". These males are also considered hybrids because they carry mtDNA of *S. pyrenaicus* [6], despite their nuclear non-hybrid genome composition that is maintained through the reproductive dynamics of the complex [reviewed in [11]].

Being composed of allopolyploid individuals, the *S. alburnoides* complex is suited for qualitative and quantitative assessments of allele-specific transcriptional control (e.g. P and A). In a recent work, Pala *et al.* [12] showed a preferential expression of A alleles and an absence of P allele transcripts in most PAA triploids from one southern population (Sorraia River, Tejo basin, additional file 1, Figure S1). Contrastingly, in two analysed northern populations (from Douro and Mondego river basins), for the majority of individuals, both C and A genome alleles were simultaneously detected, irrespective of ploidy level or genomic composition. As such, the different patterns of allele usage found within the complex correlate with the presence of P or C genomes in the hybrid triploid forms, suggesting that differential expression regulation is due to differential genome interactions [12]. Nonetheless, while for C-containing forms the specimens were collected from two distinct Northern river basins, the P-containing individuals were all from the same river (Sorraia, Tejo basin) [12,13]. Thus, this phenomenon could not be considered to be generally connected to the simultaneous presence of P and A genomes, or whether it is a population-specific feature of the Sorraia River and/or Tejo basin. This, however, is crucial information to better understand the putative genomic interactions and/or other mechanisms regulating gene transcription dynamics in this allopolyploid organism.

The overrepresentation in whole organ extracts of a specific allele could be explained by the presence of several cell types, contributing unevenly to the total RNA extracted. Moreover, this effect can be more evident in an allopolyploid context when comparing organ-specific expression patterns between individuals of different ploidy and genomic constitutions. As such, the detection of expression differences between individuals and/or between organs can be the result of mosaicism within an organ and of different levels of mosaicism between organs. Indeed, ploidy mosaicism is well established and documented in vertebrates [14,15]. Natural ploidy mosaicism appears often associated with interspecific hybridization, as in the case of the reproductive complexes of the fish *Poecilia formosa* [16], *Cobitis taenia* [17] and lizards of the genus *Lacerta* [18]. Hence, in this context, the *S. alburnoides* complex is a likely candidate for the occurrence of this phenomenon. Moreover, in some species like *Platemys platycephala* diploid-triploid mosaics appear to be geographically and population dependent [19].

To determine ploidy and gene expression profiles, we developed a flow cytometry and cell sorting protocol for *S. alburnoides* tissues. This ensured a more homogeneous cells sampling for each organ with respect to cell number, size and complexity. In these samples we determined the expression profile of three widely expressed genes (*rpl8*, *gapdh* and *β-actin*) in liver and kidney of diploid and triploid forms of *S. alburnoides* from three major Portuguese southern river basins.

Methods

(a) Specimens collection, preliminary genotyping and preparation of cell suspensions

Samples of *S. alburnoides* and *S. pyrenaicus* were collected (and handled) with the approval of the portuguese National Forest Authority (AFN, fishing credential n° 29/2011) from several locations, distributed by three major river basins, corresponding to the southern distribution range of the complex in sympatry with *S. pyrenaicus* (Tejo, Guadiana and Almargem basins) (Additional file 1, Figure S1). All individuals were brought alive to the laboratory, morphologically identified and maintained under international ethical guidelines (ASAB, 2006).

From each individual a fin clip was obtained and each specimen was identified following the method described in Morgado-Santos *et al.* [20]. DNA was obtained by standard phenol/chloroform extraction from fins and the specimens were genotyped according to Inácio *et al.* [21]. Each individual was sacrificed with an overdose of the anaesthetic MS222 and blood was collected directly from the heart, diluted in freezing solution (40 mM citric acid trisodium salt, 0.25 M sucrose, and 5% dimethyl sulfoxide) and immediately frozen at -80°C for at least 30 minutes (to allow stabilization). Liver and kidney were collected and immediately digested for 15 minutes in 0.25% Trypsin (Sigma) and mechanically dissociated/homogenized using 26 G needle syringes. A HBSS solution containing 2% FBS was added to each sample to inactivate the enzymes and an 1100 rpm centrifugation for 8 minutes at 4°C was performed. Cells were resuspended in a HBSS + 2% FBS solution and filtered through a 40 μm nylon mesh. Cell numbers, morphology and viability (percentage of living cells from each organ after digestion treatment) were assessed using a Hemocytometer and Trypan blue staining.

(b) Ploidy assessment

After preparation of the cell suspensions from liver, kidney and blood, nuclear staining was performed to assess ploidy diversity among cells of each organ in a subsample of each cell suspension. DRAQ5 (Biostatus) was added to aliquots of 0.5×10^6 or 1×10^6 cells of each cell suspension according to manufacturer instructions.

Chicken blood (2.5 pg of DNA per erythrocyte) was used as standard.

Cells were analysed on a FACSAria cytometer (BD Biosciences, San Jose, CA) equipped with both a 488 nm (15 mW output) Coherent Sapphire solid state laser (for light scatter analysis) and a 633 nm (18 mW output) JDS Uniphase HeNe air cooled laser for Draq5 excitation. Draq5 emission was detected using a 660/20 bandpass filter. Data was acquired using FACSDiva software (BD Biosciences, San Jose, CA) and acquisition of cells was performed with gating to exclude cell doublets and debris (FSC-W x FSC-A). The total number of collected events for ploidy determination was >10,000 per sample.

(c) Cell sorting

To the remaining fraction of the cell suspensions of liver and kidney (DRAQ5 free), propidium iodide (P.I.: 1/5 of stock solution at 0.5 ng/ml) was added and incubated for 20 min at room temperature. Cells were analysed on a FACSAria high-speed cell sorter using the 488 nm (15 mW output) Coherent Sapphire solid state laser for light scatter analysis and P.I. excitation. P.I. emission was detected using a 695/40 band-pass filter. Data were acquired using FACSDiva software and acquisition of cells was performed with gating to exclude cell doublets and debris (FSC-W x FSC-A), and dead cells (P.I. positive).

From the light scatter dot plots (FSC-A x SSC-A) obtained from each organ digestion, a consistent pattern of events was identified between samples of the same organ, and two main regions (A and B: A_L and B_L in liver, A_K and B_K in kidney) were defined for each organ. For a set of individuals that presented homogeneous ploidy level, one region (B_L from liver and B_K from kidney) was chosen for cell sorting to increase the intra and inter sample homogeneity. Also, from three non-mosaic individuals (Sq18, Sq29 and Sq31), composed exclusively of 3 n cells, both A and B populations from both organs were sorted to assess whether expression mosaics correlate with different cell types. In one of the individuals where ploidy mosaicism was detected, both regions (A and B) from each organ were independently sorted because they roughly corresponded to 2 n and 3 n cells.

At least 2 replicates of 100,000 cells were sorted from each organ/fish directly to Buffer RLT Plus of the AllPrep DNA/RNA Mini Kit (Qiagen) and immediately frozen at -80°C for posterior nucleic acid extraction.

(d) Genotyping and genome expression determination of the sorted cells

RNA and DNA were obtained from the previously frozen cells using AllPrep DNA/RNA Mini Kit (Qiagen).

The isolated DNA of B_L sorted cell population of each fish was used as template for the amplification of *ß-actin* gene. Genotyping of that cell population was performed based on analyses of *ß-actin* PCR products according to Sousa-Santos *et al.* [22].

From the extracted RNA, first-strand cDNA was synthesized with RevertAid First Strand cDNA Synthesis Kit (Fermentas) by using oligo dT primers. Three genes, *ß-actin*, *rpl8* and *gapdh* were amplified with specific primers (Additional file 2, Table S1) and according to the following PCR conditions: pre-heating at 94°C for 5 min, 35 cycles at 94°C for 1 min, 53°C (*rpl8*)/56°C (*gapdh* and *ß-actin*) for 1 min and 72°C for 1 min 30 s and a final extension at 72°C for 15 min. The PCR products were directly sequenced and analysed. Polymorphic sites for the two genomes (P and A) for Almargem and Guadiana fish populations were identified for the three genes using genome control sequences obtained from *S. pyrenaicus* and "nuclear non-hybrid" *S. alburnoides* from the mentioned rivers [GenBank accession numbers: JN790945; JN802520-JN802528; JN813376-JN802582]. For Tejo specimens the work of Pala *et al.* [12,13] provided the sequences for Tejo P and A genome specific polymorphisms for the three

genes [EU199435-6; EU542913-6]. In hybrid samples, the presence of cDNAs derived from single genome copies or from both genomes was determined through sequence comparison by sequence alignment using Sequencher ver. 4.0 (Gene Codes Corporation, Inc.) and based on the identified polymorphic sites between genomes (P and A). Forward and reverse sequences for each gene were obtained per individual/per organ.

Results

(a) Intra-organ differences in ploidy - Detection of mosaic individuals

A total of 40 fish were analysed using flow cytometry for ploidy determination in blood, liver and kidney cell suspensions: four *S. pyrenaicus*, three nuclear non-hybrid *S. alburnoides* and 33 hybrids *S. alburnoides* (Table 1).

All the analysed *S. pyrenaicus* and nuclear non-hybrids *S. alburnoides* displayed exclusively diploid cells in liver, kidney and blood. From the analysis of the hybrid individuals, four were identified as ploidy mosaics (Figure 1a; Table 1): three from Almargem and one from Guadiana. In all four specimens, mosaicism was detected both in liver and in kidney but not in blood (Figure 1a). For mosaic individuals, the percentage of 2n and 3n cells within each organ was assessed (Table 2). In kidney, the percentages of 2n and 3n cells were quite constant between individuals, amounting to around 50%. In liver, the inter-individual variability was higher, and in three of the four cases there were more of 3n than 2n cells composing the organ. In blood, 100% of the cells were diploid in one mosaic specimen and 100% triploid in two others. In the fourth mosaic specimen vestigial amounts, less than 1.5%, of 2n cells were detected.

Table 1 Specimens' genotype, river basin, stream of capture and ploidy status in liver, kidney and blood

Genotype[1]	Code	Basin	Stream	Liver		Kidney		Blood
				A_L	B_L	A_K	B_K	A_B
AA	Sq1	Almargem	Almargem	2n	2n	2n	2n	2n
AA	Sq22; Sq23	Guadiana	Murtega	2n	2n	2n	2n	2n
PA	Sq6[2]	Almargem	Almargem	3n	2n	2n/3n	3n	2n
PA	Sq7; Sq8	Almargem	Almargem	2n	2n	2n	2n	2n
PA	Sq24; Sq25; Sq26	Guadiana	Foupana	2n	2n	2n	2n	2n
PA	Sq27	Guadiana	Murtega	2n	2n	2n	2n	2n
PA	Sq32	Tejo	Ocreza	2n	2n	2n	2n	2n
PAA	Sq12; Sq13; Sq14; Sq15; Sq17; Sq18; Sq19; Sq20; Sq21	Almargem	Almargem	3n	3n	3n	3n	3n
PAA	Sq11[2]	Almargem	Almargem	3n	2n	2n/3n	3n	3n
PAA	Sq16[2]	Almargem	Almargem	3n	2n	3n	2n	2n/3n
PAA	Sq28; Sq29	Guadiana	Murtega	3n	3n	3n	3n	3n
PAA	Sq30[2]	Guadiana	Caia	3n	2n	3n	2n	3n
PAA	Sq31	Guadiana	Caia	3n	3n	3n	3n	3n
PAA	Sq33; Sq34; Sq35	Tejo	Ocreza	3n	3n	3n	3n	3n
PAA	Sq39; Sq40	Tejo	Sorraia	3n	3n	3n	3n	3n
PP	Sq2; Sq3; Sq4; Sq5	Almargem	Almargem	2n	2n	2n	2n	2n
PPA	Sq36; Sq37	Tejo	Ocreza	3n	3n	3n	3n	3n
PPA	Sq38	Tejo	Sorraia	3n	3n	3n	3n	3n

[1]Genotyping from DNA extracted from fin clips

[2]Ploidy mosaic specimen

A_L and B_L defined cell dot regions in liver; A_K and B_K cell dot regions in kidney

(b) Determination of genotype and allele-specific expression in mosaic and non-mosaic individuals

From the analysis of each cell suspension in the flow cytometer a light scatter dot plot (FSC-A x SSC-A) of each organ was obtained for all individuals (Figure 2). The light scatter dot plots from all blood samples presented just one homogenous population and one region was detected (A_B) (Figure 2a). From the light scatter plots obtained from liver and kidney, despite some variability found between individuals, two main dot regions, (A and B: A_L and B_L in liver, A_K and B_K in kidney) could be identified for each organ for each specimen (Figure 2b and 2c).

b1) Gene expression patterns according to organ and geographical location

The allele expression pattern of *ß-actin*, *rpl8* and *gapdh* genes of B_K and B_L cells was assessed for a total of 20 individuals pooled from the Tejo, Almargem and Guadiana samples (Table 3). As expected, all PA individuals, regardless of the basin of origin, expressed simultaneously A and P alleles (biallelic expression) in both analysed organs for the 3 analysed genes (*ß-actin*, *rpl8* and *gapdh*). In triploid PAA's from Guadiana, the expression of all 3 genes was also biallelic, both in liver and kidney. *rpl8* expression was as well consistently

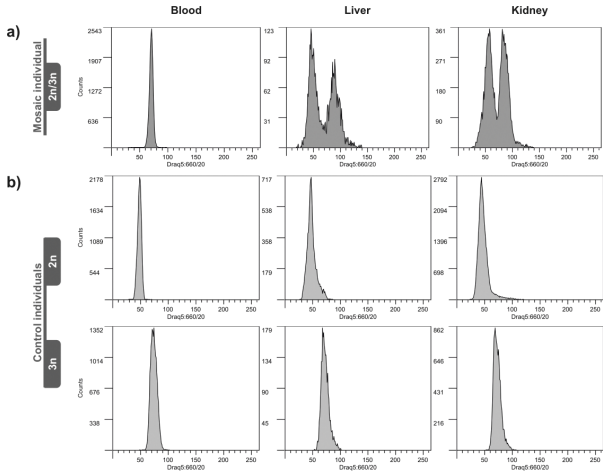


Figure 1 DNA flow histograms of *S. alburnoides* liver, kidney and whole blood cell suspensions. (a) example of one 2n/3n mosaic specimen, (b) control diploid and triploid specimens. Plots obtained from the analysis of Sq30; Sq2 and Sq9 respectively.

biallelic in both organs in all analysed triploid PAA's. On the other hand, the expression profile of *ß-actin* and *gapdh* in PAA individuals from Almargem and Tejo was more variable. Despite the majority of biallelic expression detected for the 3 genes in both organs in the individuals from Almargem, there was one individual (Sq9) where only A-*gapdh* genome transcripts were detected in kidney and in liver samples. In two individuals from Tejo (Sq33 and Sq40), only A allele expression of *gapdh* was detected in kidney, but it was biallelic in the liver of these specimens and in both organs of the other Tejo individuals. *ß-actin* expression was biallelic in liver and kidney of all individuals irrespective of the geographic origin, except in the liver of one Almargem specimen (Sq14), which presented only A transcripts.

The expression pattern of triploid PPA's from Tejo was also determined for *ß-actin*, *rpl8* and *gapdh* genes, and it was found to be biallelic for the 3 genes in both organs.

The genotype of both A and B cells in liver and kidney of controls for expression mosaic (non ploidy mosaic triploid Sq18, Sq29 and Sq31) was PAA, and the expression outcome was biallelic (Table 3) for all the

Table 2 Percentage of diploid and triploid cells in liver, kidney and blood of mosaic *S. alburnoides*

Code	Liver cells		Kidney cells		Blood	
	2n (%)	3n (%)	2n (%)	3n (%)	2n (%)	3n (%)
Sq6	31	69	56,6	43,4	100	0
Sq11	23,6	76,4	54,9	45,1	0	100
Sq16	20,6	79,4	52	48	1,3	98,7
Sq30	58,5	41,5	51,2	48,8	0	100

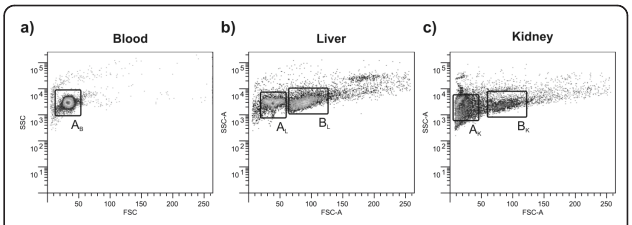


Figure 2 Typical $FSC_{vs}SSC$ dot plots obtained from the tissues digestion. (a) whole blood (b) liver and (c) kidney cell suspensions of *S. alburnoides*. Plots obtained from the analysis of Sq30.

individuals for both organs and for both A and B cell fractions.

b2) Analysis of ploidy mosaics

In two of the individuals (Sq16 and Sq30) where ploidy mosaicism was detected, the 2n and 3n cell pools (P_{3n} and P_{2n}) in liver and kidney corresponded to the light scatter defined A and B regions in each organ ($P_{3n} = A_K = A_L$ and $P_{2n} = B_K = B_L$) for both organs. This natural separation allowed sorting of 2n and 3n cells from liver and kidney without nuclear staining. The use of intercalating dyes for cellular DNA content measurements proved to be not compatible with on column DNA/RNA extraction (tested on samples Sq6 and Sq11, that were this way lost, data not shown). Only from Sq16 and Sq30 individual P_{2n} and P_{3n} sorted cells were isolated without nuclear staining but only from Sq16 good quality DNA and RNA were obtained from both diploid and triploid cell pools.

The 2n and 3n cell pools were genotyped as $P_{3n} = A_K = A_L$ = PAA genotype and $P_{2n} = B_K = B_L$ = AA genotype.

The genome specific allele expression of *gapdh, β-actin* and *rpl8* in both 2n and 3n cell pools was as well assessed. It revealed that $P_{3n} = A_K = A_L$ where P and A transcripts were detected, and in $P_{2n} = B_K = B_L$ where only A transcripts were detected.

Discussion

In the present work we studied the expression pattern of *S. alburnoides* specimens from three southern Portuguese drainages (Tejo, Guadiana and Almargem), using RNA obtained from homogeneous pools of cells form whole organs. We used flow cytometry and cell sorting to obtain homogeneous cell pools for RNA extraction and to screen for the occurrence of somatic ploidy mosaics in *S. alburnoides*.

Flow cytometry clearly revealed the occurrence of diploid-triploid mosaicism in *S. alburnoides* complex. The detected frequency of this phenomenon was approximately 10%, indicating that the diploid-triploid mosaics represent a non-regular component of the genetic system of this complex rather than a stably incorporated feature of its reproductive dynamics, as

Table 3 β-*actin, rpl8* and *gapdh* P and A allele-specific transcripts detected in liver and kidney cells of individuals from Almargem, Guadiana and Tejo populations of the *S. alburnoides* complex

Code	River Basin	River site	Ploidy	Genotype[1]	Liver expression			Kidney expression		
					β-*actin*	*rpl8*	*gapdh*	β-*actin*	*rpl8*	*gapdh*
Sq22	Guadiana	Murtega	2n	AA	A	A	A	A	A	A
Sq23	Guadiana	Murtega	2n	AA	A	A	A	A	A	A
Sq1	Almargem	Almargem	2n	AA	A	A	A	A	A	A
Sq3	Almargem	Almargem	2n	PP	P	P	P	P	P	P
Sq4	Almargem	Almargem	2n	PP	P	P	P	P	P	P
Sq5	Almargem	Almargem	2n	PP	P	P	P	P	P	P
Sq27	Guadiana	Murtega	2n	PA	PA	PA	PA	PA	PA	PA
Sq8	Almargem	Almargem	2n	PA	PA	PA	PA	PA	PA	PA
Sq29	Guadiana	Foupana	3n/3n	PAA	PA/PA	PA/PA	PA/PA	PA/PA	PA/PA	PA/PA
Sq31	Guadiana	Caia	3n/3n	PAA	PA/PA	PA/PA	PA/PA	PA/PA	PA/PA	PA/PA
Sq9	Almargem	Almargem	3n	PAA	PA	PA	A	PA	PA	A
Sq13	Almargem	Almargem	3n	PAA	PA	PA	PA	PA	PA	PA
Sq14	Almargem	Almargem	3n	PAA	A	PA	PA	PA	PA	PA
Sq18	Almargem	Almargem	3n/3n	PAA	PA/PA	PA/PA	PA/PA	PA/PA	PA/PA	PA/PA
Sq16	Almargem	Almargem	3n/2n	PAA/AA	PA/A	PA/A	PA/A	PA/A	PA/A	PA/A
Sq38	Tejo	Sorraia	3n	PPA	PA	PA	PA	PA	PA	PA
Sq40	Tejo	Sorraia	3n	PAA	PA	PA	PA	PA	PA	A
Sq36	Tejo	Ocreza	3n	PPA	PA	PA	PA	PA	PA	PA
Sq37	Tejo	Ocreza	3n	PPA	PA	PA	PA	PA	PA	PA
Sq33	Tejo	Ocreza	3n	PAA	PA	PA	PA	PA	PA	A

[1]DNA extracted from liver cells

reported in *Platemys platycephala* [19] and *Liolaemus chiliensis* [23]. Interestingly, the observed 10% ploidy variation is qualitatively different from previous reports of the same nature such as *P. formosa* [16] where this frequency was 2 orders of magnitude lower. In this case, being the occurrence of mosaic *P. formosa* very rare, the phenomenon has been considered as a mistake of a complicated reproductive system without evolutionary meaning. On another hand, being higher, the *S. alburnoides* mosaic frequency raises questions about whether the phenomenon has a real impact on the evolutionary dynamics of the species.

According to Dawley and Goddard [14], there are two possible main mechanisms that lead to diploid/triploid mosaicism: delayed fertilization and genome loss. Delayed fertilization occurs when the sperm pronucleus is slow to fuse with the female pronucleus and so, fails to participate in the first mitotic division. In this case the sperm nucleus is kept in one of the daughter cells (blastomeres) and fuses with a maternal nucleus only later, after a variable number of mitotic divisions. Consequently, a mosaic arises with triploid cells resulting from fertilization and diploid cells resulting from an initial "gynogenetic" development. This is the case of the diploid-triploid mosaics of *Misgurnus anguillicaudatus* [24] and possibly of the naturally occurring 2n/3n mosaic *P. formosa* [16]. Another mechanism is genome loss. Here, one parental chromosome set is selectively eliminated. This selective loss of a whole genomic set has been documented to occur during oogenesis of hybridogenetic unisexuals, such as *Rana esculenta* [25] and *Bufo pseudoraddei baturae* [26]. Both of the above mentioned mechanisms may be causing mosaicism in *S. alburnoides*, since this hybrid complex presents many reproductive pathways with altered oogenesis (with genomic exclusion) and spermatogenesis [reviewed in [11]].

Considering the 2n(AA)/3n(PAA) mosaic individual (Sq16), the possible routes (Figure 3) leading to this mixed genotype can be explained considering the reproductive modes of the different *S. alburnoides* forms [reviewed in [11]]. PAA triploid individuals are the most abundant form of the complex, and they are normally produced throughout the syngamy of a diploid PA oocyte with a haploid A sperm, or also by syngamy of one haploid A oocyte with a diploid PA sperm. Although uncommon, other path that leads to PAA formation is the syngamy of a diploid AA oocyte (produced by PAA females) with a haploid P sperm. If a P sperm nucleus enters a diploid AA ovum, initially remaining quiescent but later undergoing amphimixis with an early cleavage cell (Figure 3, route I), a 2n(AA)/3n(PAA) mosaic individual would arise through delayed fertilization. Another delayed fertilization scenario that could

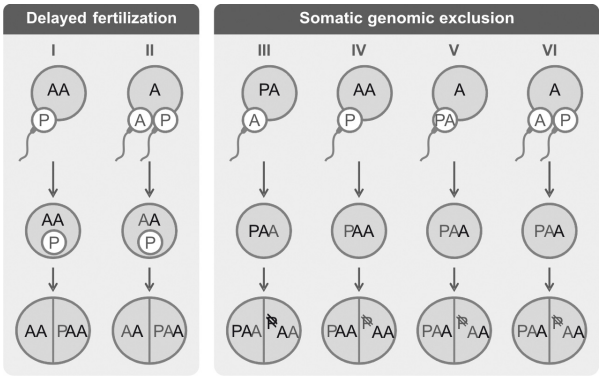


Figure 3 Possible mechanisms leading to the formation of 2n (AA)/3n(PAA) mosaic *S. alburnoides*. Alternative developmental routs within the two main mechanisms of mosaic establishment (delayed fertilization and genomic exclusion) that can lead to the formation of 2n(AA)/3n(PAA) mosaics. Eggs (large circles) and sperm (small circles with tail) contribute with P and A genomic complements (maternal genomic contributions in black and paternal genomic contributions in grey).

lead to the occurrence of a 2n(AA)/3n(PAA) mosaic is dispermy (Figure 3, route II). In such case a haploid oocyte has been fertilized by two sperm cells carrying distinct genomics sets. If karyogamy occurred only between the oocyte nucleus (A) and the sperm carrying the homologous genome, while the P sperm nucleus remains inactive during one or more embryo cleavages and only later fusing with an AA blastomere, a chimeric 2n(AA)/3n(PAA) organism would be obtained.

The 2n(AA)/3n(PAA) mosaics may also result from the loss of a whole P genome from single dividing cells in a triploid PAA embryo (Figure 3 routs III; IV, V and VI). Genomic exclusion is documented to occur during gametogenesis in hybridogenetic unisexuals, *S. alburnoides* including. Studies in the hybridogenetic water frogs *Pelophylax esculentus* [27] revealed that the genome exclusion from the germ line occurs prior to meiosis, during the prolonged phase of oogonial proliferation. So, the extension of this phenomenon to non-germinal lineages is not a big leap. In fact, the process of elimination of chromatin from pre-somatic and somatic cells is not an oddity, being in fact a very common mechanism in differentiation and development [reviewed in [3]]. The viable occurrence of 2n/3n human mosaics (or mixoploids) is also known [reviewed in [28]] and was, at least circumstantially, related to genomic exclusion and a phenomenon described as postzygotic diploidization. These human mixoploids had two paternal genomic contributions, so they originated through a process similar of what is illustrated in routes V or VI of Figure 3.

Regarding the other *S. alburnoides* specimens diagnosed as 2n/3n mosaics, we were not able to genotype the 2n and 3n cell populations from liver and kidney, so they might present other genomic compositions than 2n (AA)/3n(PAA). Therefore, the possible ways and routes that could lead to *S. alburnoides* 2n/3n mosaics may go beyond the ones sketched in Figure 3.

Another aspect worth discussing is the percentage of triploid and diploid cells that characterizes the mosaic individuals. According to Lamatsch *et al.* [16], either in the mosaics resulting from delayed fertilization or from genomic exclusion (if occurring early in development), a greater proportion of diploid cells, compared to triploid ones, would be expected. Occurring early in development, due to the lower DNA content, these diploid cells should probably replicate their DNA faster than the triploid cells and would, therefore, be able to divide more often than triploid cells. Nilsson and Cloud [29] postulated that in organs in which cells are rapidly replicating, triploid cells are prone to lose extra chromosomes and resume diploidy. So, if our results point to a phenomenon of postzygotic diploidization by genomic exclusion, it occurred in a not so early stage of development, since no strong bias was detected towards diploid cells (Table 2).

An unexpected result was found in blood ploidy measures. In this tissue, 100% of the cells were triploid in two of the mosaic specimens, 100% diploid in another and some vestigial 2n cells were detected in one sample (less than 1.5%). Some cases confirm that the use of blood is an accurate determinate of overall ploidy levels [19], once the comparison of the proportions of diploid and triploid cells in the blood with the ones determined in other tissues of the same individual, it showed only minor deviation. On the contrary, in our case, if only blood have been analysed, the mosaics would have been misdiagnosed as complete diploid and/or triploid individuals. The reasons why mosaicism is not present in the *S. alburnoides* blood samples is difficult to explain, but also in some specimens of the mosaic *P. platycephala*, blood presented a non-mosaic phenotype while some solid tissues of that same specimens were clearly 2n/3n mixoploid [19]. In one case reported in humans, a 46, XX/69,XXY mosaic also displayed a similar variation between tissues. While the 2n/3n ratio was 2:3 in fibroblasts, in blood (lymphocytes) the ratio was 24:1 [30]. An explanation for these results is that the blood is derived from the hematopoetic stem cells and has a continuous proliferating ancestry which is different to kidney and liver. While kidney and liver mosaicism may reflect a situation that goes back to the embryo when both organs were formed, the blood is reflecting the adult situation. It may well be that in the hematopoetic stem cell pool only one type of the two ploidy stages will become more prevalent. If one ploidy state is advantageous, there might be selection in the multiple rounds

of hematopoetic stem cell divisions. So 2n could be faster cycling than 3n and finally only 2n cells will be seen. On the other hand 3n stem cells might have a greater allelic repertoire and this could be advantageous.

The choice of liver, kidneys and blood as target organs was related to technical issues, because the procedure was attempted also in other organs but with no success. The analysis of gonads would have been particularly interesting because it is know from experimental crosses that triploid *S. alburnoides* females can in fact, sporadically produce haploid and triploid eggs [7].

Beyond the existence of ploidy mosaicism, also the possible occurrence of expression mosaics within the organs was cursorily prospected (Table 1: Sq18, Sq29 and Sq31). No differences were detected, neither between cell populations nor genes, being the expression pattern constantly biallelic (PA) so we have not found expression mosaicism at this level of analysis.

The prospection for mosaicism was one of the goals of this work because if happening it could have some impact in the expression patterns within and between organs. The pattern of preferential homologue genome usage previously detected for Tejo (Sorraia River population) [12,13] could have been affected or biased due to mosaicism. So, we analysed the expression pattern of three genes, *rpl8*, *gapdh* and *β-actin*, for several *S. alburnoides* individuals (which ploidy status had been assessed), not only from Sorraia River (Tejo basin), but also from some other populations of Tejo and other southern drainages (Guadiana and Almargem). We detected for all analysed specimens from Tejo a preferential biallelic expression in the cells sorted, both from liver and kidney, for *β-actin* and *rpl8* genes, and also in liver cells for *gapdh* gene. Nevertheless, P genome transcripts of *gapdh* were not detected in the kidney cells of two non-mosaic triploid PAA's, one coming from Sorraia and one from Ocreza. Consequently, we can conclude that a) the detection of only A transcripts is a phenomenon independent of ploidy mosaicism; and b) although P genomic complement is present, it is not transcribed in some tissues and from some genes, as presented and discussed by Pala *et al.* [13]. This allele silencing is not restricted to individuals from a single river (Sorraia), but also occurs in other river (Ocreza) from the same drainage (Tejo basin), and in different drainages (observed also in Almargem basin), along the range of sympatry with *S. pyrenaicus*.

When a preferential allelic usage of A in PAA fish happens, that could be interpreted as a matter of genomic homology. If genomic homology plays a role in regulating allelic expression we would predict that in PPA individuals we should detect P expression, predominantly. Therefore, we extended the analysis of Tejo triploid *S. alburnoides* to three PPA individuals, a previously not analysed genomic constitution. For these animals, expression is constantly biallelic (PA) suggesting that genetic homology is unlikely to be at play in regulating the profiles of allelic expression of triploid individuals.

Also, the occasional occurrence of ploidy mosaics does not correlate with the sporadically absent P allele expression. Only A allele expression was observed to occur in non-mosaic individuals, and when analysing the expression pattern of the Sq16 mosaic specimen (2n-AA/3n-PAA), the expression was biallelic (PA) for the 3n (PAA) cells despite the monoallelic (A) expression of the 2n (AA) cells that composed the organs of that individual.

In this work we detected the occurrence of ploidy mosaics among *S. alburnoides* specimens, but we could discard the influence of this phenomenon on the detection of variable allelic expression profiles in triploid individuals. Alternatively, as previously proposed [13], the absence of P allele transcripts in some genes of triploid PAA *S. alburnoides*, as we also report (Table 3), can be explained by the occurrence of compensation by gene-copy silencing. Consequently, PAA' triploid individuals would only transcribe two alleles per gene (PA or AA or PA'). In fact, some studies predominantly in polyploid plants [31,32] have been pointing to a process of functional diploidization as a way to balance gene dosage [33]. So, if a functional diploidization is necessary and is in fact the way through which *S. alburnoides* can cope with allopolyploidy, the ploidy status of the organism is not relevant. In this scenario, the occurrence of mixoploidy may emerge from the relaxation of the mechanisms that impose a tight control over mitosis and ploidy control.

Conclusions

We have shown that ploidy mosaicism occurs sporadically within the *S. alburnoides* complex, but in a frequency significantly higher than reported for other organisms. Moreover, we could exclude the influence of this phenomenon on the detection of variable allelic expression profiles of ubiquitously expressed genes in cells from liver and kidney of triploid individuals.

Finally, we determined that the expression patterns previously detected only in a narrow geographic range is not a local restricted phenomenon but is widespread in rivers where *S. pyrenaicus* is sympatric with *S. alburnoides*.

Altogether, our results point to interesting avenues of research on the evolutionary and mechanistic interplay between mitotic checkpoints, polyploidization and mosaicism.

Additional material

Additional file 1: Figure S1-Distribution of *S. alburnoides* in Portugal and areas of sympatry with other *Squalius* species involved in the *S. alburnoides* polyploid reproductive complex. Figure S1-Distribution of *S. alburnoides* in Portugal and areas of sympatry with other *Squalius* species involved in the *S. alburnoides* polyploidy reproductive complex. Distribution of *S. alburnoides* in Portugal and areas of sympatry with other *Squalius* species involved in the *S. alburnoides* polyploid reproductive complex. Rivers from which *S. alburnoides* and *S. pyrenaicus* were sampled are marked in red in the first panel: a) Ocreza; b) Sorraia, c) Caia; d) Murtega; e) Foupana and f) Almargem. In the second panel the major Portuguese river basins are identified.

Additional file 2: Table S1. Primer sequences and references for each gene. Table S1. Primer sequences and references for each gene. Primer sequences and references for each gene amplified for this work..

Acknowledgements
The authors thank to Miguel Morgado-Santos for his help with fishing and fish photo-identification. This work was supported by Project PTDC/BIA-BIC/110277/2009 to M.M.C. and by a grant (SFRH/BD/61217/2009) to I.M., both from Fundação para a Ciência e a Tecnologia. M.S. is supported by the Deutsche Forschungsgemeinschaft (SFB 567). E.S. and R.G. are supported by Instituto Gulbenkian de Ciência and Fundação Calouste Gulbenkian.

Author details
[1]Centro de Biologia Ambiental, Departamento de Biologia Animal, Faculdade de Ciências da Universidade de Lisboa, Campo Grande, Lisbon, 1749-016, Portugal. [2]Physiologische Chemie I, Biozentrum, University of Würzburg, Am Hubland Würzburg, 97074, Germany. [3]Instituto Gulbenkian de Ciência, Rua da Quinta Grande, Oeiras, 2780-156, Portugal. [4]Departamento de Biologia Animal, Faculdade de Ciências da Universidade de Lisboa, Campo Grande, Lisbon, 1749-016, Portugal.

Authors' contributions
MMC conceived and coordinated the study. IM was involved developing the work in all steps of the study and drafted the manuscript. ES and MS participated in the design of the study and in the critical revision of the manuscript. MPM participated in cytometry and cell sorting. RG supervised the cytometry and cell sorting assays. AI participated in fish genotyping. All authors participated in the discussion of the results and read and approved the final manuscript.

Received: 9 September 2011 Accepted: 5 December 2011
Published: 5 December 2011

References

1. Morgan TH, Sturtevant AH, Muller HJ, Bridges CB: **The mechanism of Mendelian heredity.** New York: Henry Holt and Company; 1915.
2. Székvölgyi L, Nicolas A: **From meiosis to postmeiotic events: homologous recombination is obligatory but flexible.** *FEBS J* 2010, **277(3)**:571-589.
3. Kloc M, Zagrodzinska B: **Chromatin elimination: an oddity or a common mechanism in differentiation and development?** *Differentiation* 2001, **68(2-3)**:84-91.
4. Hassold TJ: **Chromosome abnormalities in human reproductive wastage.** *Trends Genet* 1986, **2(4)**:105-110.
5. Alves MJ, Coelho MM, Collares-Pereira MJ: **Diversity in the reproductive modes of females of the *Rutilus alburnoides* complex (Teleostei, Cyprinidae): a way to avoid the genetic constraints of uniparentalism.** *Mol Biol Evol* 1998, **15(10)**:1233-1242.
6. Alves MJ, Coelho MM, Prospero MI, Collares-Pereira MJ: **Production of fertile unreduced sperm by hybrid males of the *Rutilus alburnoides* complex (Teleostei, Cyprinidae): an alternative route to genome tetraploidization in unisexuals.** *Genetics* 1999, **151(1)**:277-283.
7. Alves MJ, Gromicho M, Collares-Pereira MJ, Crespo-López E, Coelho MM: **Simultaneous production of triploid and haploid eggs by triploid *Squalius alburnoides* (Teleostei, Cyprinidae).** *J Exp Zoolog A Comp Exp Biol* 2004, **301A(7)**:552-558.
8. Crespo-López ME, Duarte T, Dowling T, Coelho MM: **Modes of reproduction of the hybridogenetic fish *Squalius alburnoides* in the Tejo and Guadiana rivers: an approach with microsatellites.** *Zoology* 2006, **109(4)**:277-286.
9. Sousa-Santos C, Collares-Pereira MJ, Almada V: **Fertile triploid males: an uncommon case among hybrid vertebrates.** *J Exp Zool A Ecol Genet Physiol* 2007, **307A(4)**:220-225.
10. Collares-Pereira MJ, Coelho MM: **Reconfirming the hybrid origin and generic status of the Iberian cyprinid complex *Squalius alburnoides*.** *J Fish Biol* 2010, **76(3)**:707-715.
11. Alves MJ, Coelho MM, Collares-Pereira MJ: **Evolution in action through hybridisation and polyploidy in an Iberian freshwater fish: a genetic review.** *Genetica* 2001, **111(1-3)**:375-385.
12. Pala I, Schartl M, Brito M, Vacas JM, Coelho MM: **Gene expression regulation and lineage evolution: the North and South tale of the hybrid polyploid *Squalius alburnoides* complex.** *Proc R Soc B Biol Sci* 2010, **277(1699)**:3519-3525.
13. Pala I, Coelho MM, Schartl M: **Dosage compensation by gene-copy silencing in a triploid hybrid fish.** *Curr Biol* 2008, **18(17)**:1344-1348.
14. Dawley RM, Goddard KA: **Diploid-triploid mosaics among unisexual hybrids of the minnows *Phoxinus eos* and *Phoxinus neogaeus*.** *Evolution* 1988, **42(4)**:649-659.
15. Doeringsfeld MR, Schlosser IJ, Elder JF, Evenson DP: **Phenotypic consequences of genetic variation in a gynogenetic complex of *Phoxinuseos-neogaeus* clonal fish (Pisces, Cyprinidae) inhabiting a heterogeneous environment.** *Evolution* 2004, **58(6)**:1261-1273.
16. Lamatsch DK, Schmid M, Schartl M: **A somatic mosaic of the gynogenetic Amazon molly.** *J Fish Biol* 2002, **60(6)**:1417-1422.
17. Janko K, Bohlen J, Lamatsch D, Flajshans M, Epplen JT, Rab P, Kotlik P, Slechtova V: **The gynogenetic reproduction of diploid and triploid hybrid spined loaches (Cobitis: Teleostei), and their ability to establish successful clonal lineages-on the evolution of polyploidy in asexual vertebrates.** *Genetica* 2007, **131(2)**:185-194.
18. Darevsky IS, Danielyan FD, Sokolova TM, Rozonov YM: **Intraclonal mating in the parthenogentic lizard species *Lacerta unisexualis*.** In *Evolution and ecology of unisexual vertebrates. Volume 466.* Edited by: Dawley RM, Bogart JP. Albany: Bulletin of New York State Museum; 1989:228-235.
19. Bickham JW, Hanks BG: **Diploid-triploid mosaicism and tissue ploidy diversity within *Platemys platycephala* from Suriname.** *Cytogenet Genome Res* 2009, **127(2-4)**:280-286.
20. Morgado-Santos M, Matos I, Vicente L, Collares-Pereira MJ: **Scaleprinting: individual identification based on scale patterns.** *J Fish Biol* 2010, **76(5)**:1228-1232.
21. Inácio A, Matos I, Machado MP, Coelho MM: **An easier method to identify the individual genomic composition of allopolyploid complexes.** *J Fish Biol* 2010, **76(8)**:1995-2001.
22. Sousa-Santos C, Robalo JI, Collares-Pereira MJ, Almada VC: **Heterozygous indels as useful tools in the reconstruction of DNA sequences and in the assessment of ploidy level and genomic constitution of hybrid organisms.** *DNA Seq* 2005, **16(6)**:462-467.
23. Lamborot M, Manzur ME, Alvarez-Sarret E: **Triploidy and mosaicism in *Liolaemus chiliensis* (Sauria, Tropiduridae).** *Genome* 2006, **49(5)**:445-453.
24. Morishima K, Oshima K, Horie S, Fujimoto T, Yamaha E, Arai K: **Clonal diploid sperm of the diploid-triploid mosaic loach, *Misgurnus anguillicaudatus* (Teleostei, Cobitidae).** *J Exp Zoolog A Comp Exp Biol* 2004, **301A(6)**:502-511.
25. Graf JD, Müller WP: **Experimental gynogenesis provides evidence of hybridogenetic reproduction in the *Rana esculenta* complex.** *Experientia* 1979, **35(12)**:1574-1576.
26. Stöck M, Lamatsch DK, Steinlein C, Epplen JT, Grosse WR, Hock R, Klapperstück T, Lampert KP, Scheer U, Schmid M, *et al*: **A bisexually reproducing all-triploid vertebrate.** *Nat Genet* 2002, **30(3)**:325-328.
27. Tunner HG, Heppich S: **Premeiotic genome exclusion during oogenesis in the common edible frog, *Rana esculenta*.** *Naturwissenschaften* 1981, **68(4)**:207-208.
28. Golubovsky MD: **Postzygotic diploidization of triploids as a source of unusual cases of mosaicism, chimerism and twinning.** *Hum Reprod* 2003, **18(2)**:236-242.
29. Nilsson EE, Cloud JG: **Extent of mosaicism in experimentally produced diploid-triploid chimeric trout.** *J Exp Zool* 1993, **266(1)**:47-50.

30. Dewald G, Alvarez MN, Cloutier MD, Kelalis PP, Gordon H: **A diploid-triploid human mosaic with cytogenetic evidence of double fertilization.** *Clin Genet* 1975, **8(2)**:149-160.
31. Paterson AH, Bowers JE, Chapman BA: **Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics.** *Proc Natl Acad Sci USA* 2004, **101(26)**:9903-9908.
32. Wang XY, Shi XL, Hao BL, Ge S, Luo JC: **Duplication and DNA segmental loss in the rice genome: implications for diploidization.** *New Phytol* 2005, **165(3)**:937-946.
33. Birchler JA, Veitia RA: **The gene balance hypothesis: from classical genetics to modern genomics.** *Plant Cell* 2007, **19(2)**:395-402.

doi:10.1186/1471-2156-12-101
Cite this article as: Matos *et al.*: Ploidy mosaicism and allele-specific gene expression differences in the allopolyploid *Squalius alburnoides*. *BMC Genetics* 2011 **12**:101.

CHAPTER 2

Supplementary data

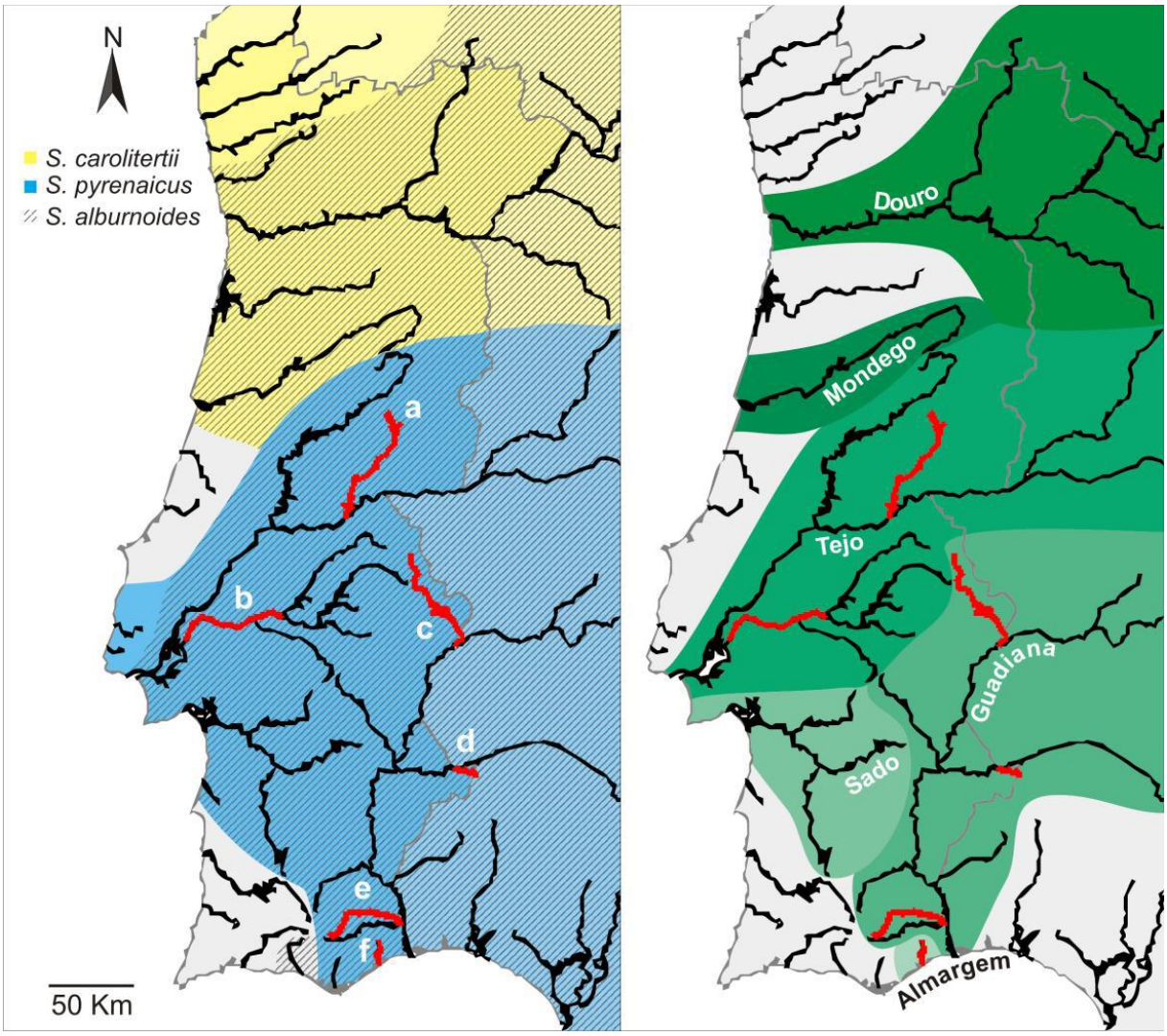


Figure S1- Distribution of *S. alburnoides* in Portugal and areas of sympatry with other Squalius species involved in the *S. alburnoides* polyploid reproductive complex.
Rivers from which *S. alburnoides* and *S. pyrenaicus* were sampled are marked in red in the first panel: a) Ocreza; b) Sorraia, c) Caia; d) Murtega; e) Foupana and f) Almargem. In the second panel the major Portuguese river basins are identified.

Table S1. Primer sequence and references for each gene

Gene	Primer	Sequence	References
β-actin	β-ACTIN-F1	5'-CAACGGCTCCGGCATGTG-3'	Pala et al., 2008
	β-ACTIN-R1	5'-TGCCAGGGTACATGGTGG-3'	Pala et al., 2008
rpl8	Rpl8 forward	5'-CTCCGTCTTCAAAGCCCATGT-3'	Pala et al., 2008
	Rpl8 reverse	5'-TGTTCCTCGCAGTCTGCCAG-3'	Pala et al., 2008
gapdh	GAPDH-F1	5'-ATCAGGCATAATGGTTAAAGTTGG-3'	Pala et al., 2008
	GAPDH-Ri	5'-GGCTGGGATAATGTTCTGAC-3'	-

CHAPTER 3

Novel Method for Analysis of Allele Specific Gene Expression in Triploid *Oryzias latipes* Reveals Consistent Pattern of Allele Exclusion

Garcia TI, **Matos I**, Shen Y, et al. Novel Method for Analysis of Allele Specific Expression in Triploid Oryzias latipes Reveals Consistent Pattern of Allele Exclusion. Gong Z, ed. PLoS ONE.9(6):e100250 (2014)

OPEN ACCESS Freely available online

PLOS ONE

Novel Method for Analysis of Allele Specific Expression in Triploid *Oryzias latipes* Reveals Consistent Pattern of Allele Exclusion

Tzintzuni I. Garcia[1], Isa Matos[2,3], Yingjia Shen[1], Vagmita Pabuwal[1], Maria Manuela Coelho[3], Yuko Wakamatsu[2], Manfred Schartl[2,4], Ronald B. Walter[1]*

1 Department of Chemistry and Biochemistry, Molecular Biosciences Research Group, Texas State University, San Marcos, Texas, United States of America, **2** Physiological Chemistry, Biozentrum, University of Würzburg, Würzburg, Germany, **3** Centro de Biologia Ambiental, Faculdade de Ciências da Universidade de Lisboa, Universidade de Lisboa, Lisboa, Portugal, **4** Comprehensive Cancer Center, University Clinic Würzburg, Würzburg, Germany

Abstract

Assessing allele-specific gene expression (ASE) on a large scale continues to be a technically challenging problem. Certain biological phenomena, such as X chromosome inactivation and parental imprinting, affect ASE most drastically by completely shutting down the expression of a whole set of alleles. Other more subtle effects on ASE are likely to be much more complex and dependent on the genetic environment and are perhaps more important to understand since they may be responsible for a significant amount of biological diversity. Tools to assess ASE in a diploid biological system are becoming more reliable. Non-diploid systems are, however, not uncommon. In humans full or partial polyploid states are regularly found in both healthy (meiotic cells, polynucleated cell types) and diseased tissues (trisomies, non-disjunction events, cancerous tissues). In this work we have studied ASE in the medaka fish model system. We have developed a method for determining ASE in polyploid organisms from RNAseq data and we have implemented this method in a software tool set. As a biological model system we have used nuclear transplantation to experimentally produce artificial triploid medaka composed of three different haplomes. We measured ASE in RNA isolated from the livers of two adult, triploid medaka fish that showed a high degree of similarity. The majority of genes examined (82%) shared expression more or less evenly among the three alleles in both triploids. The rest of the genes (18%) displayed a wide range of ASE levels. Interestingly the majority of genes (78%) displayed generally consistent ASE levels in both triploid individuals. A large contingent of these genes had the same allele entirely suppressed in both triploids. When viewed in a chromosomal context, it is revealed that these genes are from large sections of 4 chromosomes and may be indicative of some broad scale suppression of gene expression.

Citation: Garcia TI, Matos I, Shen Y, Pabuwal V, Coelho MM, et al. (2014) Novel Method for Analysis of Allele Specific Expression in Triploid *Oryzias latipes* Reveals Consistent Pattern of Allele Exclusion. PLoS ONE 9(6): e100250. doi:10.1371/journal.pone.0100250

Editor: Zhiyuan Gong, National University of Singapore, Singapore

Received February 21, 2014; **Accepted** May 22, 2014; **Published** June 19, 2014

Funding: Funding was provided partly by the NIH, ORIP, Division of Comparative Medicine (http://dpcpsi.nih.gov/orip/CM/index) grants R24-OD-011120 and R24-OD-011199. IM was supported by the Graduate Training programme 1048 of the Deutsche Forschungsgemeinschaft (http://www.dfg.de), and by the International Graduate School for Life Sciences of the University of Würzburg (http://www.graduateschools.uni-wuerzburg.de/life_sciences). Support was also provided by the Fundação para a Ciência e Tecnologia PhD grant SFRH/BD/61217/2009. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: RWalter@txstate.edu

Introduction

Allele specific expression (ASE) is an important component of gene regulation that is not well studied, but is thought to account for a major part of the phenotypic variation within and among species [1,2]. Among plants in general, and particularly in many food crops, polyploidy also plays a major role in enhancing phenotypic variation and is often associated with increased vigor and the gain of desirable traits [3]. In plants made polyploid through hybridization, homoeologous genes (ancestrally homologous genes incorporated in an allopolyploid organism) can have uneven allele specific expression levels or overall gene expression levels that differ greatly from the parents [4]. These homoeologous genes bring together their accompanying regulatory elements which interact with the rest of the regulatory machinery upon hybridization to unevenly affect allele expression and may lead to extensively altered phenotypes [5]. In order to understand the impact of allopolyploidization on a molecular genetic level, it will be necessary to study ASE on a genome-wide scale.

In addition to better understanding of plants important to our food supply, understanding ASE in polyploid states is important to human health. In many cases cancerous cells contain multiple extra chromosomes leading to partial or full polyploidy [6]. Nondisjunction events also result in partially duplicated chromosomes and are mostly incompatible with life in humans, but in other cases lead to large phenotypic disruptions [7]. All of these situations are related to the more basic question of how ASE is affected by the elevation of a diploid genome to a polyploid state.

In such a situation it may be that alleles of each gene are expressed at the same levels in the polyploid environment as in the diploid such that the total gene expression is greater than that of a parent. Alternately there could be some form of dosage

compensation such as silencing of individual alleles or of an entire haplome (genetically distinct set of chromosomes). There may be a bias for one haplome, or it could be random. In order to sufficiently answer these questions it is necessary to study the whole genome where all alleles can be distinguished, but this has proven to be problematic. It is a problem common to many polyploid biological systems and has been faced before largely in plant studies but also some animals [5–9].

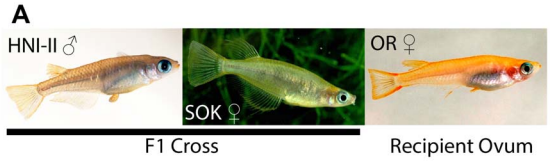
Several studies have addressed the basic issue of ASE in polyploid systems in plants, and a recent review provides an excellent summary of these findings [5]. In general a wide array of expression patterns have been observed in hybrid systems [5,8,9]. In some cases expression from each allele may be additive. For other genes, however, the expression is dominated by one allele. These effects may be tissue-specific, responsive to environmental cues, or they may be biased toward one parent. In general these observations have been made on small gene sets. High throughput studies include some use of microarrays [10,11] and a recent analysis of allopolyploid (AD) cotton using RNAseq [4].

Examples of polyploid animals are rare especially among vertebrates, but several examples exist among amphibians and teleost fish. An especially notable example is that of the *Squalius alburnoides* complex. This is a naturally occurring inter-generic hybrid population descended from *Squalius pyrenaicus* females and males of an extinct species in which individuals may contain between 2–4 genome copies [12,13]. A mechanism of dosage compensation is employed by this fish is the silencing of specific alleles, but not of an entire haplome [12].

The ability to measure ASE in these situations is vital to better understand global mechanisms of allele dosage compensation and to disentangle gene interaction networks. Thus, we have generated a model in which three haploid genomes with sufficient genetic diversity to allow determination of allele specific gene expression were combined to experimentally produce triploids. We used the small laboratory fish medaka (*Oryzias latipes*) to produce two triploid individuals through nuclear transfer of a diploid F_1 nucleus to a recipient ovum, thus incorporating haplomes from three disparate medaka strains. We then developed a computational methodology to derive allele-specific expression values from RNAseq data obtained from isolated liver RNA from the triploid and parental diploid fish. In the triploid medaka, we find that alleles are expressed at similar levels in most cases, but allele suppression is not uncommon and occurs consistently in the two triploid fish. In some cases, the suppressed alleles are completely silenced and in these cases the silenced allele is almost never derived from the maternal genome component stemming from the mother of the F_1.

Results

In order to produce artificial triploid fish, diploid hybrid F_1 embryos were first produced through the natural mating of two genetically different strains (the Houiken Niigata-II, or HNI-II, strain originates from a wild population in northern Japan and the Sokcho, or SOK, strain originates from a wild population in east Korea). Then diploid hybrid nuclei from the developing embryonic blastomeres were transplanted into unfertilized eggs of a third medaka strain (the orange-red or OR strain is derived from a commercially available variety originated from a southern Japanese wild population). The resulting triploid embryos have three genetically distinct sets of chromosomes (haplomes). Through this technique two triploid fish were produced for this study (from now on referred to as trpA and trpB) that incorporated genetic material from three divergent strains of medaka



B

reference ID	pos.	HNI	OR	SOK	Identifies
seq001	47	A/T	T	T	None
seq001	343	C	C	G	SOK
seq001	998	N	G	C	None
Discriminatory SNPs		25,014	22,678	35,481	

Figure 1. Parent strains and variant calling. (A) Parent strains and gender of donor genomes (images provided by MS). HNI-II males were mated with SOK females to produce F_1 embryos. At the blastula stage, cells were separated and diploid nuclei from them were injected into OR ova where they fused with the haploid nucleus of the oocyte. (B) Examples of variants called by VarScan. Only variant positions in which at least one strain was completely and unambiguously different from the others could be used (variant at position 343). Variant positions were unsuitable where one strain was only partially different (i.e. from heterozygosity) or where a strain had insufficient coverage to confidently call a set of observed nucleotides (variants at positions 47 and 998). The total number of dSNPs identified for each strain is indicated in the bottom line of red text.
doi:10.1371/journal.pone.0100250.g001

(Figure 1A). Both individuals were phenotypically female, and developed as apparently healthy adults, but were infertile.

Determination of dSNPs

To distinguish alleles and determine the contribution of each allele to overall gene expression we focused on using SNPs identified in the three parental strains that had been used to produce the triploids. Similar approaches for diploid organisms have been successful [14–16]. In diploid organisms SNPs can be used to discriminate expression levels of two alleles, but in the triploid case it is uncommon for a single SNP position to discriminate all three alleles. Instead we identified SNPs that were found to have only one possible nucleotide in one strain that did not overlap with the observed nucleotide possibilities in the other two strains at the same position in the same transcript.

Short read data for the three parental strains were aligned using the STAR aligner [17] (https://code.google.com/p/rna-star/). The STAR aligner was utilized since it is fast and can accurately handle several mismatches, indels and/or splice junctions. Since the reference sequences were transcripts, the ability to align over splice junctions was disabled, but alignment over short indels was still allowed. Short read samples had between of 54 to 61% of the reads alignedalign (Table 1). The reference sequences in ENSMBL arewere largely based on high throughput sequencing of the Hd-rR strain of medaka which, like the OR strain, also originates from the southern Japanese population. This was further supplemented with data from the HNI strain that is similar to the HNI-II strain and also originates from northern Japan. Therefore it is nottherefore surprisingsurprise these two strains had the most reads aligned and fewest dSNPs and indels detected. The SOK strain, originating from south Korea has the fewest reads aligned and most dSNPs and indels. Even so, using the STAR short read aligner that can work around mismatches and indels, the number of SOK reads aligned is not drastically different from the number of OR and HNI-II reads aligned (Table 1). We detected 9,913 putative indels throughout the full set of transcripts. The indels in

Table 1. Short read data and variants detected per sample.

Sample	Gender	Pre-filter fragments	Post-filter fragments	Aligned Fragments	Percent Aligned	dSNPs	indels
HNI-II	M	7.50E+07	7.20E+07	4.07E+07	57%	25,014	5,772
OR	M	6.30E+07	6.00E+07	3.68E+07	61%	22,678	3,337
SOK	M	6.10E+07	5.90E+07	3.19E+07	54%	35,481	6,085
TrpA	F	2.10E+08	2.00E+08	1.11E+08	56%	n/a	n/a
TrpB	F	2.00E+08	1.90E+08	1.13E+08	60%	n/a	n/a
synthetic	n/a	n/a	[†]1.11E+08	1.11E+08	100%	n/a	n/a

Short read data for the five samples was run in three lanes on the Illumina HiSeq 2000 platform resulting in roughly 200 million read pairs per lanelane (data available from sequence read archive associated with BioProject accession: PRJNA246137). During our filtration process it is possible some reads lose their mate while in cases where the reads overlap significantly, they are merged into one read. Therefore it is more useful to use the term fragments (the sum of pairs and single reads) as opposed to reads or pairs. In the range of 54–61% of fragments aligned successfully to our transcriptome reference sequences. The synthetic data set having been generated exclusively from the reference sequences aligned completely.

[†]The synthetic reads were not filtered, this number is the total number of synthetic reads generated.

doi:10.1371/journal.pone.0100250.t001

general did not seem to be very large (typically between 2–7 bp and up to 21 bp).

Using the set of cDNA records from the ENSEMBL 65 medaka genome annotation as our reference, we initially detected 250,982 single nucleotide variant positions where at least one of the three parent strains differed from the reference sequence. These positions however included many that did not provide useful information for the following reasons: one or more strains are heterozygous such that none can be completely distinguished from the other two (Figure 1B, position 47), the coverage in one strain is too low to make a confident call of observed sequence (annotated by VarScan as an N; Figure 1B position 998), or all three strains agree with one another but disagree with the reference. Excluding these cases left 109,581 informational sites which can distinguish the expression of one allele from the other two. For the sake of brevity, and convenience, we call these sites discriminatory SNPs (dSNPs) throughout this report. In order to measure ASE values for all three alleles in a gene, the dSNP number was further reduced to a final total of 83,173 dSNPs that occurred in transcripts containing at least one dSNP representative of each parental strain (Figure 1B discriminatory SNPs). The reference sequences we used were primarily based on data from the OR strain, and thus this strain had the fewest dSNPs detected.

Another departure from normal diploid determination of ASE is that each parental strain had a different set of dSNPs that distinguish it, and therefore we cannot simply count the reads attributable to each parental haplome. This is largely due to the wide variability in the number of dSNPs for each parental strain for some transcripts. For example, in a given reference transcript one haplome may be represented by only 1 or 2 dSNPs while another may have 20; this situation would bias read counts toward the haplome having 20 dSNPs. A second reason for not performing standard read counting is the large variability in coverage depth possible over the length of a transcript. Each dSNP can only sample the expression signal from one haplome at one position along the transcript. Thus, it may be misleading should the position of the dSNP coincide with a very low or very high depth of coverage for a given transcript.

Our strategy was to use the depth of coverage information at dSNP positions to identify the fraction of that coverage depth that was attributable to the haplome for which each dSNP was specific. Then the coverage fractions attributable to each haplome were integrated to give a single ASE value for that haplome. In order to derive ASE values from these single positions, fractional expression values are combined from dSNPs in the same gene. We were therefore constrained to the 4,282 transcripts (of 24,662 annotated in ENSEMBL version 65) that had at least one dSNP representative of each of the three haplomes. The final set of 83,173 dSNPs are present in these 4,282 transcripts resulting in an average of 19.4 dSNPs per transcript.

Determination of ASE from Normalized Coverage Depth

Short reads aligned to a reference transcriptome generally result in uneven depth of coverage (Figure 2A raw coverage). Many factors may contribute to this phenomenon and it is commonly observed in all RNAseq data. For example the expression of alternative splice forms, where two expressed splice forms may provide the common exons and thus more reads in an additive fashion, while differently incorporated exons would only be expressed at the level of each splice form. Additionally, annotated genes often include only some of the splice forms or possible exons that actually make up a locus thus making an accurate measure of gene expression more difficult. Another factor that may affect coverage variability in some transcripts stems from the inability of

short read aligners to distinguish between equally good alignment locations. Some subsequences of the reference can be very common and thus give the short read aligner a difficult choice. In these situations it is common to simply not report alignments to these regions, to randomly select one, or to report alignments to all regions. Thus, choices made in these areas can result in more or less reads aligned to them. To circumvent this problem in our analyses we chose to assess the fractional expression at dSNP sites in each transcript (Figure 2A fractional expression at dSNPs). Herein the fractional expression values were multiplied by the geometric mean of coverage depth of the transcript in which they occur in order to obtain an allele specific expression value for each dSNP position. We chose the geometric mean of coverage depth as a measure of gene expression in order to mitigate the effects of transcript length and lessen the bias in apparent expression value due to spikes in coverage. The dSNP expression values specific to each strain are then averaged in each transcript to arrive at an allele specific expression (ASE) value for each allele in the transcript (Figure 2A fractional expression at dSNPs). The ASE values calculated may be found in Table S3.

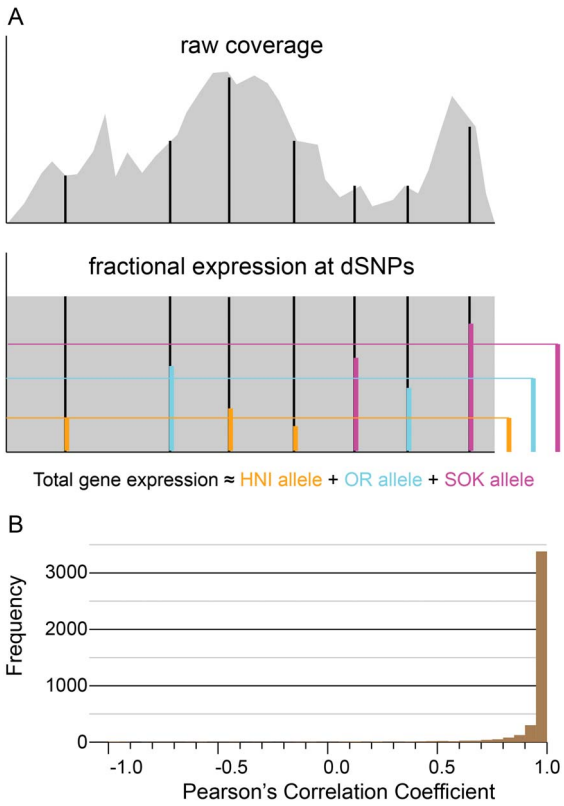


Figure 2. Illustration of ASE method and analysis of artificial data. (A) Cartoon of the raw coverage observed when RNAseq reads are mapped to an example mRNA transcript and the effect of normalizing coverage across the transcript. In both raw and normalized coverage plots, black vertical lines indicate the positions of dSNPs. In the cartoon of normalized coverage, the colored vertical bars at dSNP positions indicate the contribution of the discernible allele to the overall expression, and the colored bars to the right indicate the average of expression values measured at dSNP sites for each allele. (B) Correlation of known ASE values to calculated ASE values for the synthetic data test. The calculated allele-specific expression values are compared with the original known values for each transcript by calculating a Pearson's correlation coefficient. This will measure how well the trend in the calculated ASE values matches the trend in the original ASE values. Over 75% of the transcripts (2,628) had a correlation coefficient greater than 0.8.
doi:10.1371/journal.pone.0100250.g002

Synthetic Data Test

To evaluate the ability of this technique to accurately establish allele specific gene expression values, we devised a set of synthetic data in which these values were known. We generated an artificial set of short read data from the full set of medaka transcripts. To generate these synthetic reads, we produced a strain-specific set of transcript sequences wherein the strain-specific SNP nucleotides were substituted into the reference sequence. The total number of reads generated for each transcript was the same as that found by mapping short reads from trpA to the reference sequences, and we kept the allele balances consistent with those measured in trpA. For 91% of the transcripts in this data set (3,891 out of 4,282), we obtained a correlation of greater than 0.8 when comparing measured to actual allele expression values, with the bulk of these (3,684, 86%) having a correlation coefficient of 0.90 or greater (Figure 2B).

We ruled out several factors that were speculated to adversely affect the analysis in these 9% of cases including overall expression, numbers of dSNPs per transcript, and frequently found sub-sequences. The details of our efforts are given in Text S1. The randomness of generated reads likely had a role in the poor correlation of some transcripts in cases where the three alleles had known values very close to one another. Presume for example the known values were 255, 260, and 265 for the three alleles. These quantities of reads would have been generated from random locations along the length of the strain-specific references to represent each allele. Some noise therefore is introduced and the reconstructed ASE values may have been 11, 12, and 11. A reasonably close result which nevertheless results in a correlation coefficient of 0.

Validation of SNP Calls and ASGE Trends in the Triploid Fish

We selected seven genes with extreme allelic expression patterns for validation (Table S1 and Figure S1). We performed Sanger sequencing of the PCR products for each target gene in the triploid fish and for the three parental strains. All but two of the 36 SNP calls at dSNP sites were validated in all seven sequenced transcripts (Table S2). The trends in usage of specific alleles in triploid fish was also assessed using a previously published method [12,18]. The trends in allele-specific expression observed by Sanger sequencing were also consistent with our RNAseq-based ASE method (Figure 3 and Table S2).

The Geometric Mean of Coverage Depth Compares Well to Read Counting

We now discuss exclusively RNAseq reads from the biological samples aligned to the set of 4,282 ASE-compatible medaka transcripts. We are using the geometric mean of coverage in each transcript as a measure of the total expression of that transcript. Each dSNP gives us the fractional expression of one allele measured from the coverage depth at that position. These fractional expression values are averaged over all dSNPs specific to each strain in order to arrive at a final ASE value for the three strains. We observe a strong correlation between the sum of calculated ASE values and the geometric mean of the coverage in each transcript (Figure 4A). This indicates the method employed accurately divides the overall expression between the three alleles.

In order to examine whether the geometric mean of coverage depth was a good measure of gene expression, we compared it to

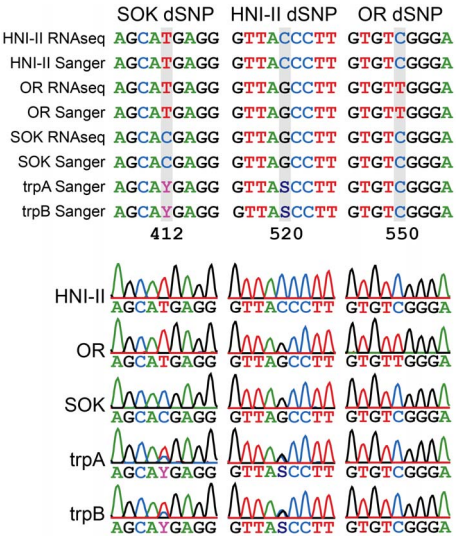


Figure 3. An example of parental dSNP and tri-hybrid ASGE validation. The represented transcript is ENSORLT00000013099 (ensembl Transcript ID). One dSNP per parental genome is represented. (A) RNA-seq and Sanger obtained sequence alignments between the parental strains (SOK, HNI-II and OR) and the Sanger obtained sequences for both trpA and trpB triploids. (B) Aligned chromatograms of the three parental strains and trpA and trpB. In both triploids the allele expression pattern determined by Sanger is consistent with the one obtained by RNA-Seq for this transcript. In this example SOK and HNI-II dSNPs are observed in the triploid, but the OR dSNP is not observed.
doi:10.1371/journal.pone.0100250.g003

the commonly used method of counting fragments mapped to each transcript. The Spearman's correlation was calculated comparing the whole gene expression values calculated by the two methods. A very high correlation is observed between the two triploids (trpA $r = 0.941$, trpB $r = 0.938$ see Figure 4B). We chose to use the Spearman's correlation because, as shown in Figure 4B, the rate of increase in the normalized expression values is reduced as higher fragment counts are reached. This is likely due to the normalized expression value being insensitive to the length of the original transcripts whereas the read counting method produces higher values for longer transcripts.

A Major Set of Allele-imbalanced Genes have Silenced One Allele

Since we only have two individuals in our analysis we selected two arbitrary boundaries in order to help interpret the data. We are interested in dividing genes into groups that may be differentially expressed in the triploid environment from those that seem to be expressed at similar levels to the average of parental expression. To this end we select a threshold of 2-fold change either up or down in gene expression for each triploid with respect to the average parental gene expression. The coefficient of variation (c_v) is used here as a normalized measure of dispersion of the allele specific expression values. A low c_v indicates the three alleles for a given gene have relatively equal expression to one another. As the c_v rises the expression of the three alleles becomes more divergent. The highest possible c_v is 1.73 which corresponds to the condition where one allele shows some expression while the other two alleles are totally shut down. The data indicate a significant clustering of genes have a c_v of $\sim$0.86. This corresponds to the situation in which one allele is not expressed (zero or near-zero expression levels) while the other two alleles make up the bulk

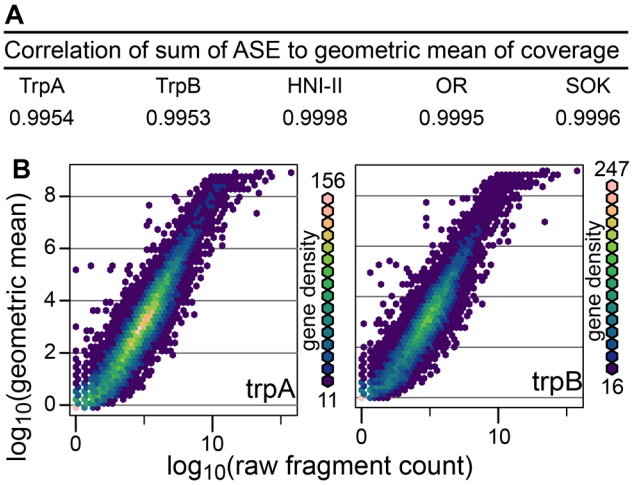


Figure 4. Comparison of calculated ASE values to whole gene expression. (A) Correlation of sum of ASE values per transcript (calculated only from dSNP sites) to geometric mean of each gene (calculated from coverage over entire transcript). (B) Plot of geometric mean of read coverage against raw fragment count showing a strong correlation between the two.
doi:10.1371/journal.pone.0100250.g004

of expression. We use this as a second arbitrary threshold to divide our data. These boundaries are shown in four quadrants of plots in Figure 5 A and B. Quadrant 0 represents transcripts in which ASE values are relatively equal to one another and in which the overall gene expression is similar to that of the parental strains (3,166 transcripts in trpA and 3,415 transcripts in trpB with 2,918 shared between them). Quadrant I represents transcripts in which ASE values are more highly dispersed but in which the overall expression values are similar to the average of the parental strains (468 transcripts in trpA and 440 transcripts in trpB with 309 shared between them). Quadrant II represents transcripts in which overall expression differs from the parents but in which the ASE values are similar to one another (533 transcripts in trpA and 338 transcripts in trpB with 149 transcripts shared between them). Quadrant III holds transcripts with dispersed ASE values and in which overall expression has changed from parental strains (115 transcripts in trpA and 89 transcripts in trpB with 38 shared).

The distribution of coefficients of variance has a main peak ($c_v \approx 0.2$) corresponding to transcripts with alleles that express at roughly equal levels in a triploid individuals (Figure S1). There is also a second peak ($c_v \approx 0.86$) that corresponds to transcripts in which one allele does not appear to be expressed (Figure S1).

ASE Imbalance Categories are Similar in both Triploid Fish

Allele specific expression values for each transcript were analyzed to identify allele expression imbalances that may indicate exceptionally high or low expression of one allele with respect to the other two. We first applied a goodness of fit test where the null hypothesis is that alleles express equally. This resulted in 1,593 transcripts in trpA and 1,447 transcripts in trpB for which the resultant p-value was less than 0.01. These sets of transcripts which deviate from equal allele expression were then broken down into groups indicative of exceptionally high or low expression of one allele.

In order to identify high or low expressing alleles, we chose to use the median of allele expression in each transcript as a basis for comparison since it is more resistant to outliers than the median. This is especially true in cases where the number of observations is

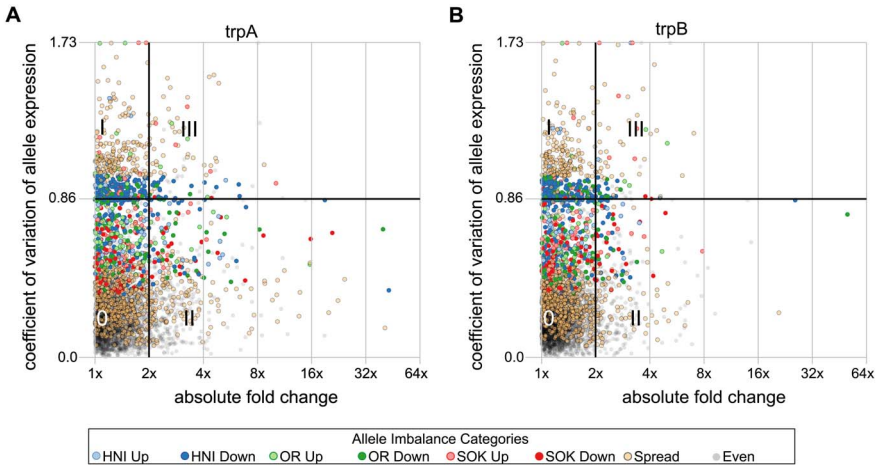


Figure 5. Overall gene expression compared to the dispersion in allele specific expression. A comparison of the change in gene expression with the dispersion of allele specific expression in trpA and trpB are plotted in panels (A) and (B) respectively. The horizontal axes indicate differential expression of the triploids with respect to the parent fish lines. The average of whole gene expression in trpA and trpB is compared to the average of the 3 parent species. The vertical axis of (A) indicates the coefficient of variation (c_v) of ASE values for each transcript in trpA, while (B) indicates the same quantity for trpB. c_v values near 0 indicate that the three alleles are expressed at near equal levels, and increasing c_v values indicate a greater dispersion of allele-specific expression. A line at a c_v value of 0.86 is drawn because this value correlates with the situation where one allele is shut down entirely and the other two are expressed at similar levels to one another. Similarly the line at a c_v value of 1.73 correlates with transcripts in which two alleles have been shut off. Additionally, a grid is drawn to separate the plot into four quadrants. Quadrant '0' has transcripts in which ASE values are least dispersed and gene expression is similar to the average of parental strains. Quadrant I contains transcripts which are expressed at levels similar to those in the parents but in which ASE levels are more highly dispersed. Quadrant II contains transcripts in which expression levels are dissimilar to the parents, but with low ASE dispersion. Quadrant III contains transcripts in which expression levels were dissimilar to the parents and ASE dispersion was increased. Points are colored to reflect categories defined by the balance of allele expression in each transcript (allele imbalance categories).
doi:10.1371/journal.pone.0100250.g005

small and a single outlier will have a very strong effect on the mean. We selected an arbitrary boundary of 2-fold above or below the median as a threshold for selection as a high or low expressing allele. Six separate categories indicate one of each of the three alleles were expressed at 2-fold above, or below the median expression of the three alleles in a given gene (these categories are: HNI Up, HNI Down, OR Up, OR Down, SOK Up, and SOK Down). Another 'spread' category indicates genes have deviated from the expectation of equal expression, but not been incuded into another category, and the final 'even' category contains transcripts for which deviation from equal expression was not rejected (legend of Figures 5 and 6).

These groups help to identify trends in the distribution of c_v. In both trpA and trpB the second peak ($c_v \approx 0.86$) is primarily composed of transcripts in the HNI-Down category followed by those in the OR-Down category (Figure S2). Of the 4,282 transcripts analyzed, 358 were found to have one allele suppressed 2-fold below the median in both triploids, while 104 were found to have an allele expressed 2-fold higher than the median allele expression in both triploids. When genes in these categories are displayed in their chromosomal context, strong similarities appear between the two triploid fish (Figure 6). In fact 3,353 transcripts are in the same category in both triploids.

Discussion

Allele-specific Suppression is a Common Feature

Overall, most genes seem to be expressed at similar total levels in triploid fish as those in the parental strains. In the case of an imbalance the data suggest it is more common for one allele to be suppressed than it is for one allele to dominate expression of the gene (Figure 5 and Figure S2). A similar feature in which certain genes are effectively diploidized has been observed in triploid individuals of the naturally occurring intergenic hybrid *S. alburnoides* complex [12,13,18]. There does not seem to be a global preference for any given allele in transcripts based on either differential gene expression or allele specific gene expression. On the other hand, in both triploids, a majority of transcripts wherein one allele has been suppressed to expression levels at or near zero (Figure 5 and Figure S2 coefficient of variation near 0.87) are in only two categories: HNI down and OR down (Figure S2). Whatever mechanism is active in this regulation, it appears as though it has a bias for the donor female-derived haplome. In both triploids the male-derived haplome (HNI-II) is the most drastically suppressed and it is followed by the haplome of the recipient ovum (SOK). On the contrary, the donor female-derived haplome (OR) was nearly unaffected by drastic allele suppression. This could be reflective of some form of allele suppression that favors the donor maternal genome. This may be somehow associated with the different packing states of the chromatin since the diploid nucleus taken from the blastula cell had been in an active state whereas the chromatin of the recipient ovum may have still been in a more dormant and packed form. Thus it may be that the donor female haplome was able to impose some allele silencing regime on the male haplome and the recipient female haplome.

ASE Imbalances are Similar between Triploid Fish

ASE analysis of two experimentally produced medaka triploids shows a consistent pattern of allele-specific gene regulation between the two triploid individuals. Of the 4,282 transcripts in our analysis, 3,353 are in the same ASE imbalance categories in both triploid fish (Figure 6). This implies common regulatory mechanism(s) may direct ASE in both fish. This is surprising since polyploidy is not commonly found in medaka populations, thus there is no expectation for the existence of selective pressures to develop a regulatory mechanism that would act consistently on

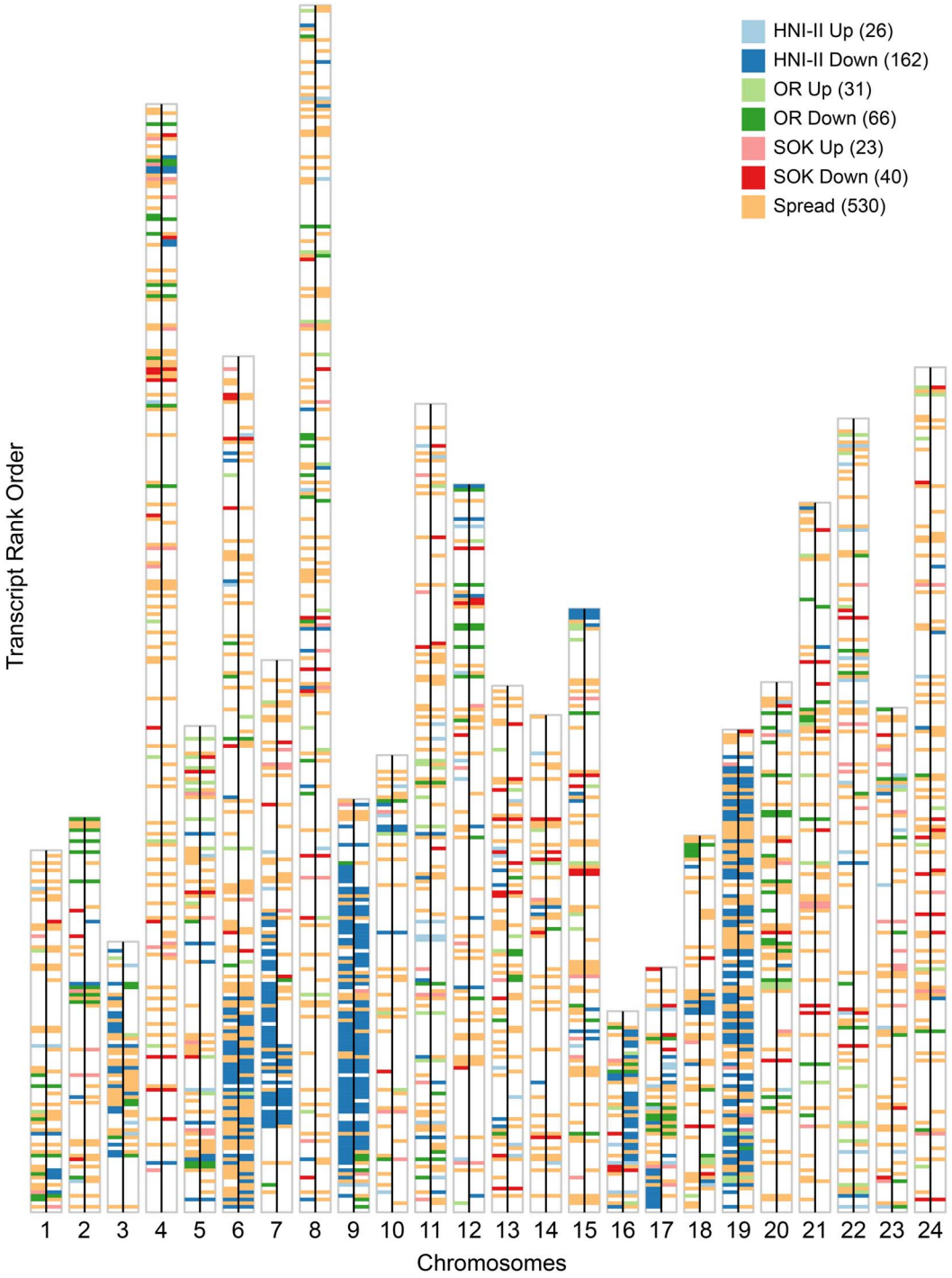


Figure 6. dSNP-complete transcripts which can be placed in chromosomes. Each chromosome is represented by large vertical white bar outlined in gray. Transcripts are represented by horizontal bars of uniform size and are placed in the rank order in which they occur in each chromosome. The bars are colored to indicate the allelic imbalance category to which the transcript belongs based on exceptional high or low expressing alleles. Blank spaces represent transcripts that could not be said to deviate from equal allele expression. Each chromosome is divided into a left and right half by a black line. The left half of each chromosome gives the plot for trpA, while the right half gives the plot for trpB. The tallest bar (that for chromosome 8) is comprised of 329 transcripts in the order they occur on the chromosome with ties being assigned sequential ranks. The total numbers of transcripts that fall into the same category in both trpA and trpB are indicated in parenthesis after each category name in the legend.

doi:10.1371/journal.pone.0100250.g006

ASE in the medaka genome in a polyploid state. In some cases in plants particular crosses have predictable patterns of ASE dominance/suppression [3] and in *S. alburnoides* the pattern of suppression is different between, yet consistent within, geographical populations [13]. In both of these cases, however, the similarities are due to preferential suppression of one whole

haplome whereas we do not see evidence of this in the medaka triploids.

Some chromosomes have large regions in which one allele is predominantly affected in a similar manner. Specifically, chromosomes 6, 7, 9 and 19 each have large regions in which most of the allelic imbalances observed resulted from suppression of the HNI allele in both triploid fish (Figure 6). We explored the possibility the four chromosomes might have been related through duplication in the teleost-specific whole genome duplication event [21]. However, according to a recent analysis, these chromosomes are thought to have originated from separate ancestral chromosomes [22], so it is unlikely that they have a common set of ancestral regulatory regions. One possible explanation is that the triploid fish are genetically very similar and so the possible regulatory schemes are likely limited. The three parental strains are genetically very homogeneous due to the closed colony breeding in which genetic bottlenecks also occur. Additionally these artificial triploid fish could share a common parentage. Donor blastula cells and host ova were pooled during the nuclear transfer proceedure and it is possible the donor nuclei for both triploids were derived from sister blastula cells leading to a very similar set of genetic material from both HNI-II and SOK strains.

Improvements Could Expand the Available Gene Set

In the current study we limited ourselves to producing ASE values for transcripts in which each allele was represented by at least one SNP. We were thus limited to a set of 4,282 transcripts out of 24,662 possible transcripts representing 4,181 of 19,687 genes that were annotated in the medaka genome (ENSEMBL version 65). One way it may be possible to expand the data set is to incorporate cases where one allele lacks dSNPs. In this case the ASE value of the remaining allele can be extrapolated from the calculated values of the two alleles with dSNPs. Doing this would expand the number of transcripts in our analysis by 1,728 or 40%.

One of the limitations on the number of transcripts amenable to ASE analysis was the variant calling step. We set very conservative thresholds when calling consensus sequences using VarScan, which likely precluded many lowly expressed transcripts from inclusion into our analysis. It is likely that many more variants could be reliably identified using an approach that specifically targets the genome and provides more even coverage, such as exome sequencing [19]. With a more robust and complete set of SNPs, it should be possible to significantly increase the number of genes in the analysis.

Another limitation of our method is that RNA-seq reads from three divergent strains were aligned to one common reference. This may cause a bias in relative expression levels since too many sequence differences occuring near one another may be a barrier to read alignment for those that do not originate from the same species or strain as the reference [20]. We ultimately used the STAR aligner which is permissive and can accurately align short reads over mismatches and indels and this aligner proved capable of greatly increasing the number of dSNPs able to used compared with others we employed (i.e., Bowtie).use. Additionally using the STAR aligner we found the balance of overall allele expression to be nearly equal (Figure S3) with only slight biases remaining against the most distant medaka strain. A more accurate yet more restrictive procedure is outlined in a recent report by Stevenson et. al. where only sites without such clusters of sequence variants are considered for ASE [20].

In summary, this complex data set has revealed several interesting biological features of the molecular-genetic activities of experimentally produced triploid medaka. The data made available by RNAseq based polyploid ASE analysis provide a highly detailed basis for the future analysis of genetic regulatory networks. This was enabled by the method we describe here for determining allele-specific expression in polyploid organisms on a large scale. Much of the software developed was created in such a way as to accommodate any ploidy number so as to be applicable to the more common diploid and/or the rare/exceptional higher-ploidy organisms with little modification. We expect this will expand our ability to understand the importance of ASE in other biologically and medically interesting systems.

Methods

Ethics Statement Regarding Animal Subjects

The research presented here complies with the applicable EU and national German legislation governing animal experimentation, especially the Tierschutzgesetz der Bundesrepublik Deutschland (German Federal Law of Animal Protection). The institution at which animal experiments were carried out is controlled by the Tierschutzbeauftragte (Animal Protection Officer) of the University of Wu?rzburg, Dr. Wolfgang Geise (Stabsstelle Arbeits-, Tier- und Umweltschutz, Marcusstraβe 9-11, D-97070 Würzburg), and therefore by the Veterinary Office of the District Government of Lower Franconia, Germany (Authorization number: 55.2-2531.01-49/08). Animal research conducted under this study has been approved by the institutional animal care and use committee of the University of Würzburg.

Parental Strains

Three strains were used as parents to generate the allotriploid fish (Figure 1A). The OR (orange-red) strain of medaka, *Oryzias latipes* [23], is derived from a commercially available orange-red variety that primarily originated from a southern Japanese wild population. The SOK (Sokcho) strain of medaka [24], originated from a wild population in east Korea. The HNI-II (Houiken-Niigata-II) strain of medaka [25] (recently taxonomically described as separate species, *Oryzias sakaizumi*) is a strain originating from a wild population of the north of Japan. All three strains were maintained as closed colony stocks and propagated in the aquarium facilities of the Biocenter in the University of Würzburg under standard conditions [26].

Donor Cells and Recipient Egg Preparation

F_1 embryos were obtained from crossing males of HNI-II with females of SOK. Donor cells were obtained from these embryos. Eggs from OR females were used as recipients. Donors and recipients were prepared according to Niwa *et al.* [27]. Briefly, 20 to 30 donor F_1 embryos at the early blastula stage were dechorionated with medaka hatching enzyme solution. Their blastoderms were dissociated into single cells. The cells were then collected by centrifugation and stored until use (up to 6 h) at 4°C in a buffer solution containing 0.25 M sucrose, 120 mM NaCl, 0.5 mM spermidine trihydrochloride (Sigma, St. Louis, MO), 0.15 mM spermine tetrahydrochloride (Sigma) and 15 mM HEPES (pH 7.3). Mature unfertilized eggs were collected from the ovary of female fish and kept in a balanced salt solution (BSS) for medaka [28] at 18°C until use (up to 5 hrs).

Nuclear Transfer

Nuclear transfer was performed according to Niwa *et al.* [27] with small modifications. An oil pressure injector made by the technical department of University of Würzburg connected to a micromanipulator (MM 33, Märzhäuser, Wetzlar, Germany) was used along with a stereomicroscope (MZ16F; Leica, Weltzlar, Germany). Also, the entire procedure was performed at 7°C. Six

days after the nuclear transplant, normally developing embryos were dechorionated with medaka hatching enzyme solution and kept at 26°C in BSS supplemented with 100 units/mL penicillin + 100 µg/mL streptomycin [29] until hatching. Hatched larvae were reared normally to the adult stage.

RNA Isolation

The liver of two triploid female medaka and the liver of one male of each parental strain (HNI-II, SOK and OR) ware collected in RNAlater (Qiagen) and used thereafter for RNA isolation. Total RNA was obtained with the RNeasy Mini Kit (Qiagen) and DNase treated on-column with the RNase-free DNase Set (Qiagen). Evaluation of integrity and quantification of the extracted RNA was performed with Nanodrop 1000 (Thermo Scientific) and 2100 Bioanalyser (Agilent Technologies) equipment. All five samples presented a RIN value of above 9 (Bioanalyser). RNA was divided in aliquots of at least 15 µg per sample and stored in RNAstable TM tubes (MoBiTec) at −80°C until further processing.

RNA Sequencing

RNAseq library build and sequencing steps were performed at Expression Analysis (Durahm, NC). Purified, poly-A selected liver RNA from each of the two triploid individuals were sequenced in one lane each of an Illumina HiSeq instrument. RNA from the three parental lines was multiplexed into a third lane, and all three lanes were sequenced as 100 bp paired ends. The resultant short reads were filtered for quality using a custom filtration pipeline. In general, steps include removing adapter sequence, trimming away low quality regions, and merging overlapping reads. Less than 5% of reads were lost during filtration and the parental strain short read libraries each consisted of approximately 64 million reads, while each of the triploid fish libraries consisted of around 200 million reads (Table 1). Short read libraries have been deposited in the sequence read archive (SRA) under BioProject accession: PRJNA246137 (https://www.ncbi.nlm.nih.gov/bioproject/).

SNP Calling

Reads from each of the three parental strains were separately aligned to the reference transcript sequences (Medaka cDNA 'all' not '*ab initio*' records from ENSEMBL v 65) using STAR version 2.3.0e (linux 64-bit pre-compiled binary) [17] to produce output files in the SAM format which were converted to sorted BAM files using samtools (Protocol S1). Each sample was aligned separately to produce three output files in total. These files were then converted to BAM format, sorted, and finally converted to mpileup format using samtools version 0.1.18 [30]. These mpileup files were then inputs for the VarScan (version 2.3.6 varscan.sourceforge.net) [31,32] mpileup2cns tool which produces an IUPAC ambiguity code for each position of the reference sequences for which the quality constraints are satisfied. We selected conservative constraints such that a position must have an overall coverage depth of 15x to be considered at all. Reads must have an read quality of 25 or more at the position of the variant to be considered. In order to call a variant, it must be supported by 20% of the reads that cover it and in lower coverage cases a minimum of 5 reads must support it. These two settings specifically should serve to minimize the likelihood of erroneous calls from sequencing errors. Further, a p-value (calculated via Fisher's Exact Test and indicates the likelihood of the call) must be below 0.01. Lastly a strand filter is applied to help identify variants that could be the result of PCR over-amplification. Full command line parameters and a flow chart describing this process are in Protocol S1.

Finding Discriminatory SNPs

The consensus calls for the three parental strains were compared and any position where one strain was found to be different from the other two strains was noted. In order for a position to be considered different in one strain, that strain had to be homozygous for the difference, and the observed nucleotides in the other strains had to be supported by the same constraints as were described in the SNP calling section above. The discriminatory SNPs (dSNPs) were then annotated with Ensembl transcript and gene IDs and grouped by transcript. Full command line parameters and a flow chart describing this process are in Protocol S1.

Determining ASE

Two main steps are required to enable the determination of ASE values. The first step involves normalizing to control for sample size variations and scaling coverage. The first step appropriately scales overall transcript expression values to enable comparisons between samples. In the second step the values of discriminating SNPs are integrated to estimate the contribution of each genome to the overall expression of a given gene.

Short reads from triploid samples were mapped using the STAR aligner, and the resulting SAM output was converted to mpileup format as previously described (Protocol S1). Then a perl script was used to extract the depth of coverage for each position in the reference sequences in each sample, and calculate the geometric mean of the depth of coverage for each transcript as a measure of the expression value. Only positions for which the coverage depth is greater than 0 are considered, therefore we define the covered length l_c to be the subset of positions in the transcript that have a coverage depth of 1 or greater. The geometric mean of the coverage depth over each transcript (g_t) with covered length (l_c) and coverage depth at each position (d_i) is described below.

$$g_t = \left(\Pi_{i=1}^{l_c} d_i\right)^{1/l_c}$$

We then use the geometric mean of coverage depth in a transcript as a measure of transcript expression.

Because the amount of data generated varies by sequencing lane and sample, we next normalized the data to control for variations in overall sample size. Briefly, the geometric mean of each transcript expression value across samples is taken as a representative transcript expression value. Then the transcript expression values in each sample are divided by their associated representative expression values. The median of these quotients in a given sample is then taken as the size factor with which to adjust each individual transcript expression value in that sample. This method is the same as is described by the authors of the DESeq package [33] with the exception that instead of total read counts here we use the geometric mean of read depth as our transcript expression value.

Next the fractional expression is measured at dSNP sites. This simply involves identifying what fraction of reads covering a dSNP position are attributed to the strain identified by that dSNP then multiplying that by the geometric mean of the coverage depth.

The fractional expression (f_i) at position i for a given strain s is given by the following formula:

$$f_i = g_t \frac{d_{i,s}}{d_i}$$

where g_t is the geometric mean of coverage depth in the transcript,

d_i is the number of reads covering position i, and $d_{i,s}$ is the number of reads covering position i that are attributable to strain s.

Only transcripts that have at least one dSNP representative of each allele are considered. The fractional expression values determined for all of the dSNPs of a given allele for a given transcript are averaged to determine that allele-specific expression (ASE) value. Only values greater than zero are considered for the signal averaging, but if all expression values are zero then that is the reported value. Full command line parameters and a flow chart describing this process are in Protocol S1.

Parental dSNP and Hybrid ASE Validation

From the list of transcripts identified as presenting informative dSNPs, seven were selected for further analysis: ENSORLT00000001009; ENSORLT00000013099; ENSORLT00000024856; ENSORLT00000008958; ENSORLT00000014111; ENSORLT00000013388; ENSORLT00000012489. The selection of these transcripts was based on the allelic expression patterns observed in the tri-hybrids. The selected targets are representative of different extreme possibilities of allelic usage in this "three allelic" context (Figure S1). Specific primers for each of these transcripts (Table S1) were designed based on the sequences alignment of the three alleles (OR, HNI-I, SOK) with Bioedit v7.2.0.

From an aliquot of each RNA sample, first-strand cDNA was synthesized with RevertAid First Strand cDNA Synthesis Kit (Fermentas). Amplification of each target transcript was performed for each sample (Table S1) according to the following PCR conditions: pre-heating at 95°C for 5 min, 35 cycles at 95°C for 30 s, 55°C or 57°C for 30 s and 72°C for 45 s and a final extension at 72°C for 10 min. The PCR products were Sanger sequenced and the sequences analyzed (Sequencher ver. 4.0, Gene Codes Corporation, Inc.) in order to validate the SNP calling between the three parental lines and the presence of expression derived from any single allele, any two, or all three alleles in the tri-hybrids.

Synthetic Data Set

The synthetic data were generated to match the calculated ASE values of trpA. The total number of reads generated per transcript were the same as found by aligning short reads from trpA to the reference. The fraction derived from each strain was determined by multiplying this total, by the fraction of total expression attributed to each strain. For example, if for a given transcript, the ASE values of HNI-II, OR, and SOK were 10, 20, and 30 respectively and the total number of fragments aligned to that transcript were 500, then the number of reads generated for HNI-II, OR, and SOK variants would be $\frac{10}{60}\times 500 = 83$, $\frac{20}{60}\times 500 = 167$, and $\frac{30}{60}\times 500 = 250$ respectively. In order to generate the fragments contributed by each strain we created strain-specific reference sequences with the strain-specific SNPs substituted in to the reference transcript sequence. Then the required number of fragments were generated as paired 100 bp reads with a fragment size of 250 bp taken at randomized start positions along the transcript length. This short read data set was then analyzed using the same software pipeline.

Supporting Information

Figure S1 Genes selected for dSNP confirmation. Genes selected for confirmation of dSNPs were taken from all 4 quadrants and are shown here as red circles overlaid on a background of grey points showing the change in gene expression vs. the dispersion of allele specific expression in trpA. The horizontal axis indicates differential expression of the triploids with respect to the parent fish lines. The average of whole gene expression in trpA and trpB is compared to the average of the 3 parent species. The vertical axis of indicates the coefficient of variation (c_v) of ASE values for each transcript in trpA. c_v values near 0 indicate that the three alleles are expressed at near equal levels, and increasing c_v values indicate a greater dispersion of allele-specific expression.
(TIF)

Figure S2 Stacked histograms of c_v values in all allelic imbalance categories. Stacked histograms of coefficient of variation of allele expression values in transcripts grouped by allele imbalance categories. A c_v value near 0.87 is consistent with complete suppression of one allele. This shows the clear preference for HNI-II and OR silencing (spike in bin of c_v values covering 0.85 to 0.90).
(TIF)

Table S1 Primers for dSNP validation. Primers used for validation of dSNPs by Sanger sequencing are listed along with the gene and transcript IDs and other descriptive information from ENSEMBL version 65. The quadrants listed are a reference to those defined in Figure 6.
(XLSX)

Table S2 Confirmation of dSNPs by Sanger sequencing. This table lists the Sanger sequencing results of dSNP sites from seven transcripts and whether or not they confirm the consensus nucleotide calls made by VarScan using RNAseq data. In total 32 out of 36 dSNPs are confirmed. The four misses were shown to be heterozygous in the parental strains by Sanger sequencing.
(XLSX)

Table S3 Allele specific expression values. This table lists allele specific expression values for trpA and trpB and whole gene expression for the three parental line samples as calculated by our methods. The non-bold text in blue, green, and red colored cells lists the expression values detected for HNI-II, OR, and SOK alleles respectively for trpA and trpB. The bold text lists whole gene expression values for trpA, trpB, HNI-II, OR, and SOK as calculated by our software pipeline. The whole gene expression values are the sum of allele-specific expression for each transcript. ENSEMBL transcript IDs are used to identify each transcript.
(TXT)

Protocol S1 Flowcharts and command lines for software tools used. An extensive set of flow charts with command line options used for our analysis. This includes several custom perl scripts which are available upon request.
(PDF)

Text S1 Tests to determine association of several factors with low accuracy of ASE. A short summary of extra tests done to investigate possible causes of low accuracy of ASE measurements.
(DOCX)

Acknowledgments

We would like to thank the reviewers for their insights.

Author Contributions

Conceived and designed the experiments: MS IM YW TIG RW MMC. Performed the experiments: IM YW TIG. Analyzed the data: TIG IM. Wrote the paper: TIG IM YW. Contributed technical expertise to design of computational analysis tools: YS VP.

References

1. Johnson NA, Porter AH (2000) Rapid Speciation via Parallel, Directional Selection on Regulatory Genetic Pathways. J Theor Biol 205: 527–542. Available: http://www.sciencedirect.com/science/article/pii/S0022519300920708. Accessed 2013 Nov 8.
2. Levine M (2002) How insects lose their limbs. Nature 415: 848–849. Available: http://dx.doi.org/10.1038/415848a. Accessed 2013 Nov 8.
3. Doyle JJ, Flagel LE, Paterson AH, Rapp RA, Soltis DE, et al. (2008) Evolutionary genetics of genome merger and doubling in plants. Annu Rev Genet 42: 443–461. Available: http://www.ncbi.nlm.nih.gov/pubmed/18983261. Accessed 2013 Dec 17.
4. Yoo M-J, Szadkowski E, Wendel JF (2013) Homoeolog expression bias and expression level dominance in allopolyploid cotton. Heredity (Edinb) 110: 171–180. Available: http://www.ncbi.nlm.nih.gov/pubmed/23169565. Accessed 2013 Aug 13.
5. Madlung A (2012) Polyploidy and its effect on evolutionary success: old questions revisited with new tools. Heredity (Edinb). Available: http://www.nature.com.libproxy.uthscsa.edu/hdy/journal/vaop/ncurrent/full/hdy201279a.html. Accessed 2012 Nov 14.
6. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, et al. (2010) The landscape of somatic copy-number alteration across human cancers. Nature 463: 899–905. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 2826709&tool = pmcentrez&rendertype = abstract. Accessed 2013 Aug 6.
7. McFadden DE, Kwong LC, Yam IY, Langlois S (1993) Parental origin of triploidy in human fetuses: evidence for genomic imprinting. Hum Genet 92: 465–469. Available: http://www.ncbi.nlm.nih.gov/pubmed/7902318. Accessed 2013 Aug 13.
8. Adams KL, Cronn R, Percifield R, Wendel JF (2003) Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. Proc Natl Acad Sci U S A 100: 4649–4654. Available: http://www.pnas.org/content/100/8/4649.full. Accessed 2013 Dec 17.
9. Mochida K, Yamazaki Y, Ogihara Y (2003) Discrimination of homoeologous gene expression in hexaploid wheat by SNP analysis of contigs grouped from a large number of expressed sequence tags. Mol Genet Genomics 270: 371–377. Available: http://www.ncbi.nlm.nih.gov/pubmed/14595557. Accessed 2013 Dec 17.
10. Rapp RA, Udall JA, Wendel JF (2009) Genomic expression dominance in allopolyploids. BMC Biol 7: 18. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 2684529&tool = pmcentrez&rendertype = abstract. Accessed 2013 Dec 12.
11. Inácio A, Pinho J, Pereira PM, Comai L, Coelho MM (2012) Global analysis of the small RNA transcriptome in different ploidies and genomic combinations of a vertebrate complex–the Squalius alburnoides. PLoS One 7: e41158. Available: http://dx.plos.org/10.1371/journal.pone.0041158. Accessed 2012 Dec 7.
12. Pala I, Coelho MM, Schartl M (2008) Dosage compensation by gene-copy silencing in a triploid hybrid fish. Curr Biol 18: 1344–1348. Available: http://www.ncbi.nlm.nih.gov/pubmed/18771921. Accessed 2012 Nov 7.
13. Pala I, Schartl M, Brito M, Malta Vacas J, Coelho MM (2010) Gene expression regulation and lineage evolution: the North and South tale of the hybrid polyploid Squalius alburnoides complex. Proc Biol Sci 277: 3519–3525. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 2982235&tool = pmcentrez&rendertype = abstract. Accessed 2013 Dec 4.
14. Skelly DA, Johansson M, Madeoy J, Wakefield J, Akey JM (2011) A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. Genome Res 21: 1728–1737. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 3202289&tool = pmcentrez&rendertype = abstract. Accessed 2012 Oct 31.
15. Tang F, Barbacioru C, Nordman E, Bao S, Lee C, et al. (2011) Deterministic and stochastic allele specific gene expression in single mouse blastomeres. PLoS One 6: e21208. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 3121735&tool = pmcentrez&rendertype = abstract. Accessed 2013 Aug 8.
16. Zhai R, Feng Y, Zhan X, Shen X, Wu W, et al. (2013) Identification of transcriptome SNPs for assessing allele-specific gene expression in a super-hybrid rice Xieyou9308. PLoS One 8: e60668. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 3629204&tool = pmcentrez&rendertype = abstract. Accessed 2013 Aug 13.
17. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, et al. (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29: 15–21. Available: http://bioinformatics.oxfordjournals.org/content/29/1/15. Accessed 2014 Mar 19.
18. Matos I, Sucena E, Machado MP, Gardner R, Inácio A, et al. (2011) Ploidy mosaicism and allele-specific gene expression differences in the allopolyploid Squalius alburnoides. BMC Genet 12: 101. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 3276436&tool = pmcentrez&rendertype = abstract. Accessed 2013 Aug 26.
19. Meynert AM, Bicknell LS, Hurles ME, Jackson AP, Taylor MS (2013) Quantifying single nucleotide variant detection sensitivity in exome sequencing. BMC Bioinformatics 14: 195. Available: http://www.biomedcentral.com/1471-2105/14/195. Accessed 2013 Dec 3.
20. Stevenson KR, Coolon JD, Wittkopp PJ (2013) Sources of bias in measures of allele-specific expression derived from RNA-seq data aligned to a single reference genome. BMC Genomics 14: 536. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 3751238&tool = pmcentrez&rendertype = abstract. Accessed 2013 Dec 11.
21. Santini F, Harmon LJ, Carnevale G, Alfaro ME (2009) Did genome duplication drive the origin of teleosts? A comparative study of diversification in ray-finned fishes. BMC Evol Biol 9: 194. Available: http://www.biomedcentral.com/1471-2148/9/194. Accessed 2013 Dec 14.
22. Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, et al. (2007) The medaka draft genome and insights into vertebrate genome evolution. Nature 447: 714–719. Available: http://www.ncbi.nlm.nih.gov/pubmed/17554307. Accessed 2010 Jul 21.
23. Bubenshchikova E, Ju B, Pristyazhnyuk I, Niwa K, Kaftanovskaya E, et al. (2005) Generation of fertile and diploid fish, medaka (Oryzias latipes), from nuclear transplantation of blastula and four-somite-stage embryonic cells into nonenucleated unfertilized eggs. Cloning Stem Cells 7: 255–264. Available: http://www.ncbi.nlm.nih.gov/pubmed/16390261. Accessed 2013 Dec 17.
24. Sakaizumi M, Jeon SR (1987) Two divergent groups in the wild populations of medaka Oryzias latipes (Pisces: Oryziatidae) in Korea. Korean J Limnol: 13–20.
25. Hyodo-Taguchi Y (1996) Inbred strains of the medaka, Oryzias latipes. Fish Biol J Medaka: 11–14.
26. Nanda I, Hornung U, Kondo M, Schmid M, Schartl M (2003) Common spontaneous sex-reversed XX males of the medaka Oryzias latipes. Genetics 163: 245–251. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 1462404&tool = pmcentrez&rendertype = abstract. Accessed 2013 Aug 14.
27. Niwa K, Ladygina T, Kinoshita M, Ozato K, Wakamatsu Y (1999) Transplantation of blastula nuclei to non-enucleated eggs in the medaka, Oryzias latipes. Dev Growth Differ 41: 163–172. Available: http://www.ncbi.nlm.nih.gov/pubmed/10223712. Accessed 2013 Aug 14.
28. Iwamatsu T (1983) A new technique for dechorionation and observations on the development of naked eggs in Oryzias latipes. J Exp Zool: 83–89.
29. Wakamatsu Y, Ozato K, Hashimoto H, Kinoshita M, Sakaguchi M, et al. (1993) Generation of germ-line chimeras in medaka (Oryzias latipes). Mol Mar Biol Biotechnol: 325–332.
30. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 2723002&tool = pmcentrez&rendertype = abstract. Accessed 2013 Aug 7.
31. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, et al. (2009) VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics 25: 2283–2285. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid = 2734323&tool = pmcentrez&rendertype = abstract. Accessed 2013 Aug 14.
32. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, et al. (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 22: 568–576. Available: http://genome.cshlp.org/content/22/3/568. Accessed 2013 Aug 14.
33. Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11: R106. Available: http://www.ncbi.nlm.nih.gov/pubmed/20979621. Accessed 2011 Jul 18.

CHAPTER 3

Supplementary data

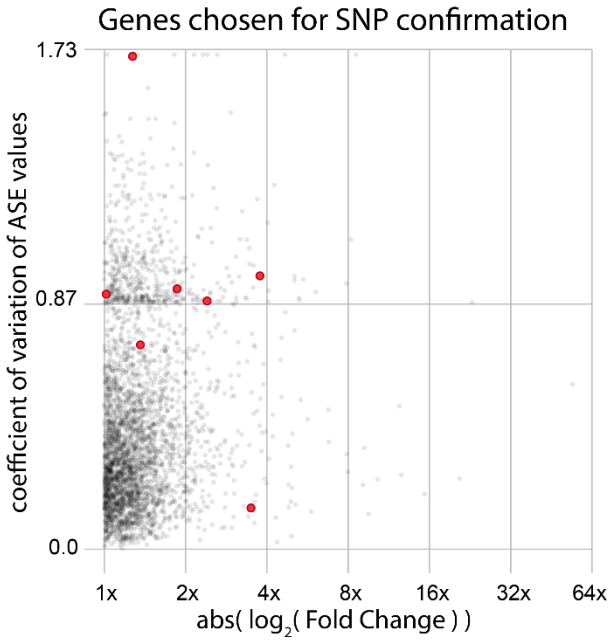


Figure S1. Genes selected for dSNP confirmation. Genes selected for confirmation of dSNPs were taken from all 4 quadrants and are shown here as red circles overlaid on a background of grey points showing the change in gene expression vs. the dispersion of allele specific expression in trpA. The horizontal axis indicates differential expression of the triploids with respect to the parent fish lines. The average of whole gene expression in trpA and trpB is compared to the average of the 3 parent species. The vertical axis of indicates the coefficient of variation (cv) of ASE values for each transcript in trpA. cv values near 0 indicate that the three alleles are expressed at near equal levels and increasing cv values indicate a greater dispersion of allele-specific expression.

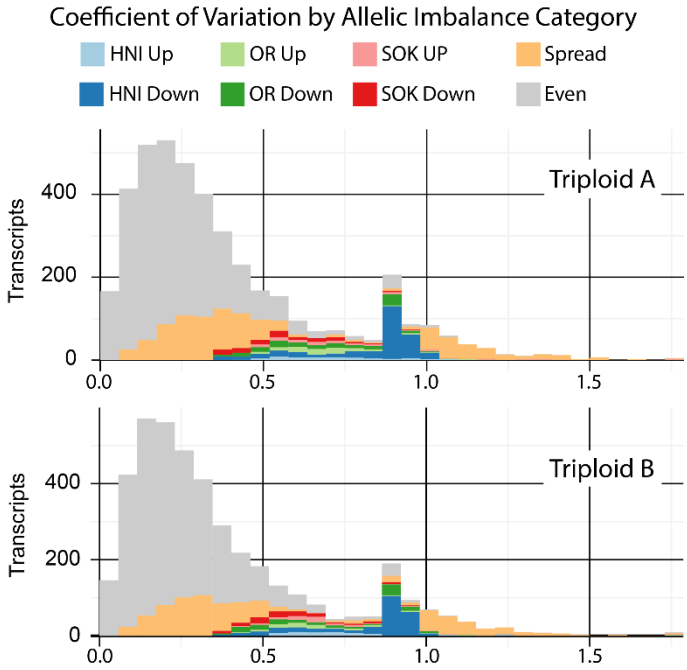


Figure S2. Stacked histograms of cv values in all allelic imbalance categories. Stacked histograms of coefficient of variation of allele expression values in transcripts grouped by allele imbalance categories. A cv value near 0.87 is consistent with complete suppression of one allele. This shows the clear preference for HNI-II and OR silencing (spike in bin of cv values covering 0.85 to 0.90).

Table S1. Primers for dSNP validation. Primers used for validation of dSNPs by Sanger sequencing are listed along with the gene IDs and other descriptive information from ENSEMBL version 65. The quadrants listed are a reference to those defined in Figure 6.

ENST	Quadrant	EXT_ Name	Description	Primers (5'-3')	TA
ENSORLT00000001009	III	OLA.4344-201	-	PF- GGATGTGAACGGGAAGGAT PR-GGCTGAGGAGCTTCTTGATG	57°c
ENSORLT00000013099	III	ANXA3	annexin A3 [Source:HGNC Symbol;Acc:541]	PF-TCTGCAGGAGAGCATTGAAA PR-GTTGGTCAGTCAGCATCCAA	55°c
ENSORLT00000024856	III	MT-CO2	mitochondrially encoded cytochrome c oxidase II [Source:HGNC Symbol;Acc:7421]	PF-GATGCAGCTTCACCCGTTAT PR-CGGTATACTCATAACTTCAATACCAC	55°c
ENSORLT00000008958	I	PXMP2	peroxisomal membrane protein 2, 22kDa [Source:HGNC Symbol;Acc:9716]	PF-CAAAATTGGAACCCCAGCTA PR-CCTCCCATCCTTTAGCTTCC	55°c
ENSORLT00000014111	I	EIF4EBP1 (2 of 2)	eukaryotic translation initiation factor 4E binding protein 1 [Source:HGNC Symbol;Acc:3288]	PF-CACCACGAGCCTGGAGAT PR-GGGAGGTTATTGGGAGGTGT	55°c
ENSORLT00000013388	0	A0FDJ6_ORYLA	60S ribosomal protein L8 [Source:RefSeq peptide;Acc:NP_001098379]	PF-TCAAGGGGATTGTGAAGGAC PR-CAGCAACCACACCGACAAC	57°c
ENSORLT00000012489	I	-	-	PF-GATCGTGTCGAGCACAAAGA PR-TCGTTCACACAAAGTGGATCA	57°c

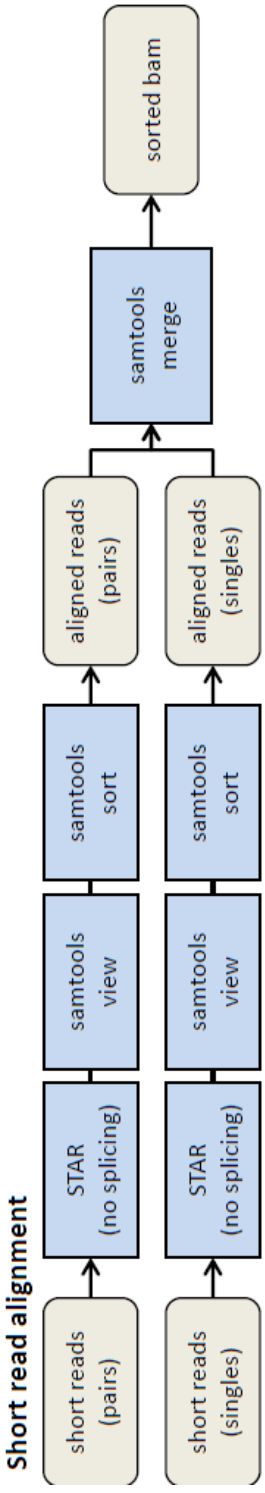
Table S2. Confirmation of dSNPs by Sanger sequencing. This table lists the Sanger sequencing results of dSNP sites from seven transcripts and whether or not they confirm the consensus nucleotide calls made by VarScan using RNAseq data. In total 34 out of 36 dSNPs are confirmed. The four misses were shown to be heterozygous in the parental strains by Sanger sequencing.

ENST	EXT_Name	Sanger Sequencing Call						HNI-II	OR	SOK	Status
		SNP Position	HNI-II	OR	SOK	Tri-hyb A	Tri-hyb B	RNA-Seq SNP Call			
ENSORLT00000001009	OLA.4344-201	249	A	A	T	W	W	A	A	T	confirmed
		327	C	T	C	Y	Y	C	T	C	confirmed
		342	T	C	C	C	C	T	C	C	confirmed
ENSORLT00000013099	ANXA3	412	T	T	C	Y	Y	T	T	C	confirmed
		439	A	C	C	M	M	A	C	C	confirmed
		440	C	A	A	M	M	C	A	A	confirmed
		520	C	G	G	S	S	C	G	G	confirmed
		550	C	T	C	C	C	C	T	C	confirmed
		582	G	T	T	K	K	G	T	T	confirmed
		583	C	G	G	S	S	C	G	G	confirmed
ENSORLT00000024856	MT-CO2	274	T/c	C	C	C	C	T	C	C	confirmed
		276	A	A	G	A	A	A	A	G	confirmed
		294	T	T	C	T	T	T	T	C	confirmed
		303	A/g	G	G	G	G	A	G	G	confirmed
ENSORLT00000008958	PXMP2	330	C	T	C	Y	Y	C	T	C	confirmed
		365	A	C	C	C	C	A	C	C	confirmed
		386	C	C	G	S	S	C	C	G	confirmed
		411	C	C	T	Y	Y	C	C	T	confirmed
ENSORLT00000014111	EIF4EBP1 (2 of 2)	108	A	G	G	R	R	A	G	G	confirmed
		111	A	G	A	R	R	A	G	A	confirmed
		123	C	C	T	C	C	C	C	T	confirmed
ENSORLT00000013388	A0FDJ6_ORYLA	253	T	T	C	Y	Y	T	T	C	confirmed
		296	C	C	T	Y	Y	C	C	T	confirmed
		347	G	A/r	G	R	R	G	A	G	miss
		377	A	C	C	M	M	A	C	C	confirmed
		491	C	C	T	Y	Y	C	C	T	confirmed
		497	C	A	A	M	M	C	A	A	confirmed
		500	A	G	A	R	R	A	G	A	confirmed
ENSORLT00000012489		1547	T	T	G	K	K	T	T	G	confirmed
		1555	C	C	A	M	M	C	C	A	confirmed
		1579	A	C	C	M	M	A	C	C	confirmed
		1721	G	G	A	R	R	G	G	A	confirmed
		1746	T	A	A	W	W	T	A	A	confirmed
		1753	G	T/g	G	K	G	G	T	G	miss
		1804	A	A	G	R	R	A	A	G	confirmed
		1827	C	C	T	Y	Y	C	C	T	confirmed

Table S3. Allele specific expression values. This table lists allele specific expression values for trpA and trpB and whole gene expression for the three parental line samples as calculated by our methods. The non-bold text in blue, green, and red colored cells lists the expression values detected for HNI-II, OR, and SOK alleles respectively for trpA and trpB. The bold text lists whole gene expression values for trpA, trpB, HNI-II, OR, and SOK as calculated by our software pipeline. The whole gene expression values are the sum of allele-specific expression for each transcript. ENSEMBL transcript IDs are used to identify each transcript.

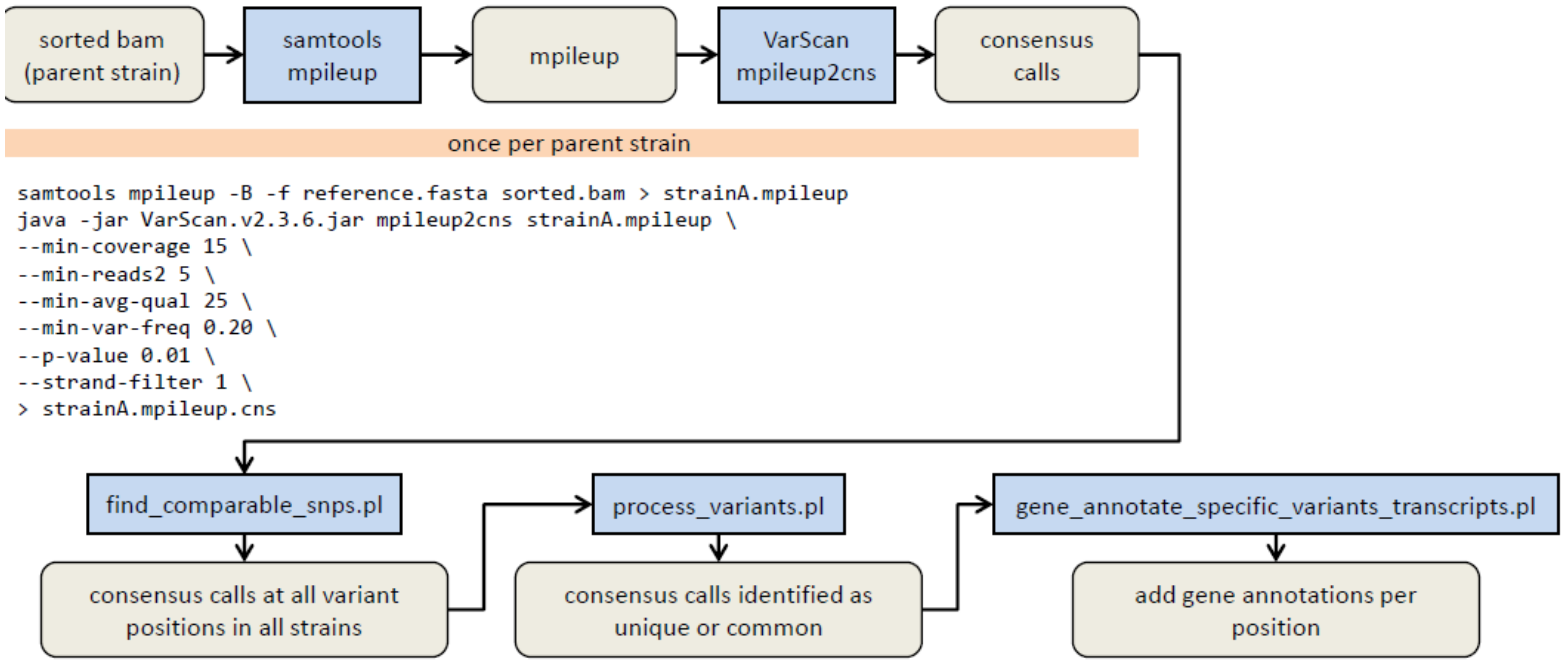
Assessible at:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4063754/bin/pone.0100250.s005.txt

Protocol S1. Flowcharts and command lines for software tools used. An extensive set of flow charts with command line options used for our analysis. This includes several custom perl scripts which are available upon request.




```
## Align paired reads
STAR \
--runMode alignReads \
--genomeLoad LoadAndKeep \
--genomeDir ~/stardb/0.1atipes_cDNA \
--readFilesIn /path/to/reads/reads_1.fastq /path/to/reads/reads_2.fastq \
--runThreadN $THREADS \
--outFileNamePrefix paired/ \
--outFilterMismatchNoverLmax 0.05 \
--outFilterMultimapNmax 10  \
--alignIntronMax 20 \ # these two together will
--alignIntronMin 21 \ # disable splicing in STAR
--outSAMattributes All \
--outStd SAM | \
samtools view -S -h -b -u - | samtools sort -@ $THREADS - paired/paired.sorted

## Align single reads
STAR \
--runMode alignReads \
--genomeLoad LoadAndKeep \
--genomeDir ~/stardb/0.1atipes_cDNA \
--readFilesIn /path/to/reads/reads_S.fastq \
--runThreadN $THREADS \
--outFileNamePrefix single/ \
--outFilterMismatchNoverLmax 0.05 \
--outFilterMultimapNmax 10  \
--alignIntronMax 20 \ # these two together will
--alignIntronMin 21 \ # disable splicing in STAR
--outSAMattributes All \
--outStd SAM | \
samtools view -S -h -b -u - | samtools sort -@ $THREADS - single/single.sorted

## Merge sorted aligned reads into one bam file
samtools merge merged.sorted.bam paired/paired.sorted.bam single/single.sorted.bam
```

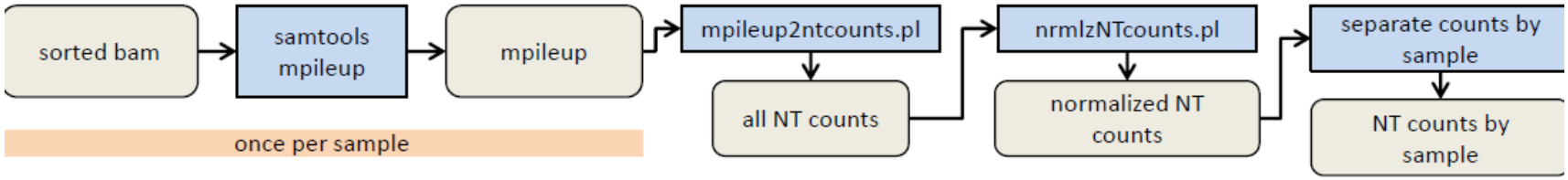
Identify SNPs




```
samtools mpileup -B -f reference.fasta sorted.bam > strainA.mpileup
java -jar VarScan.v2.3.6.jar mpileup2cns strainA.mpileup \
--min-coverage 15 \
--min-reads2 5 \
--min-avg-qual 25 \
--min-var-freq 0.20 \
--p-value 0.01 \
--strand-filter 1 \
> strainA.mpileup.cns
```

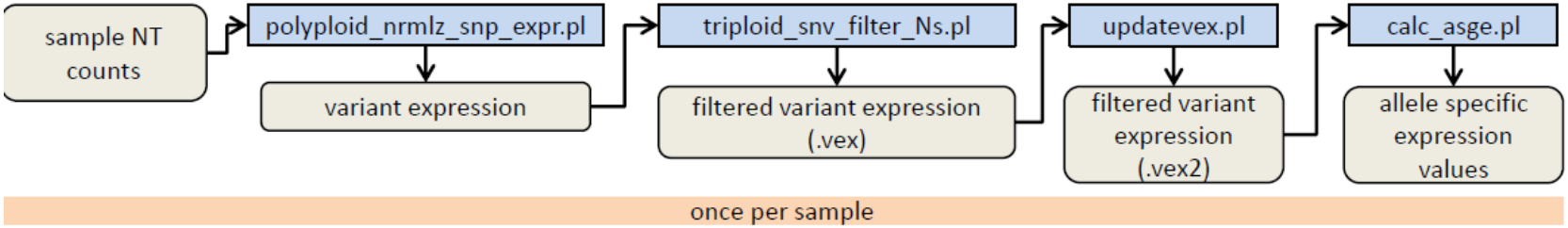
```
# Identify variants in all strains and consensus calls (as IUPAC ambiguity codes) at the same position in all parental strains
find_comparable_snps.pl reference.fasta \
strainA strainA.mpileup.cns \
strainB strainB.mpileup.cns \
strainC strainC.mpileup.cns \
> strainA_strainB_strainC.comp.cns

# Determine whether variants are uniquely found in one or more strains or whether the observed nucleotide at a given
# position and in a given strain are also seen in that position in other strains.
process_variants.pl --consensus-file=strainA_strainB_strainC.comp.cns > strainA_strainB_strainC.specific.variants

# Add annotations identifying ENSEMBL gene ID from ENSEMBL gene annotation in GTF format
gene_annotate_specific_variants_transcripts.pl genes.gtf strainA_strainB_strainC.specific.variants \
> strainA_strainB_strainC.ENSG.specific.variants
```

Normalize for sample size




```
# Once for each sample: use samtools to generate an mpileup file from the sorted bam
samtools mpileup -B -f refseq.fasta sample.sorted.bam > sample.mpileup

# Count occurrences of each nucleotide at each position in all samples and report as large tabular matrix
mpileup2ntcounts.pl strainA.mpileup strainB.mpileup strainC.mpileup \
triploidA.mpileup triploidB.mpileup triploidC.mpileup > all.ntcounts

# Normalize counts for sample size
nrmlzNTCounts.pl all.ntcounts > all_nrml_ntcounts2

#Separate normalized counts by sample
perl -e '$a=`head -n 1 all_nrml_ntcounts2`;
chomp $a;
@b=split/\t/,$a;
$n=(@b-2)/4-1;
for $i (0..$n) {
            $s=$i*4+3;
            $e=$s+3;
            $name=$b[$s-1];
            $name=~s/A\.(\S+)/$1/;
            @args=("cut -f ".join(",",1,2,$s..$e)." all_nrml_ntcounts2 > $name.nrml.ntcounts");
            system @args;
}'
```

Integrate normalized NT counts to calculate ASE

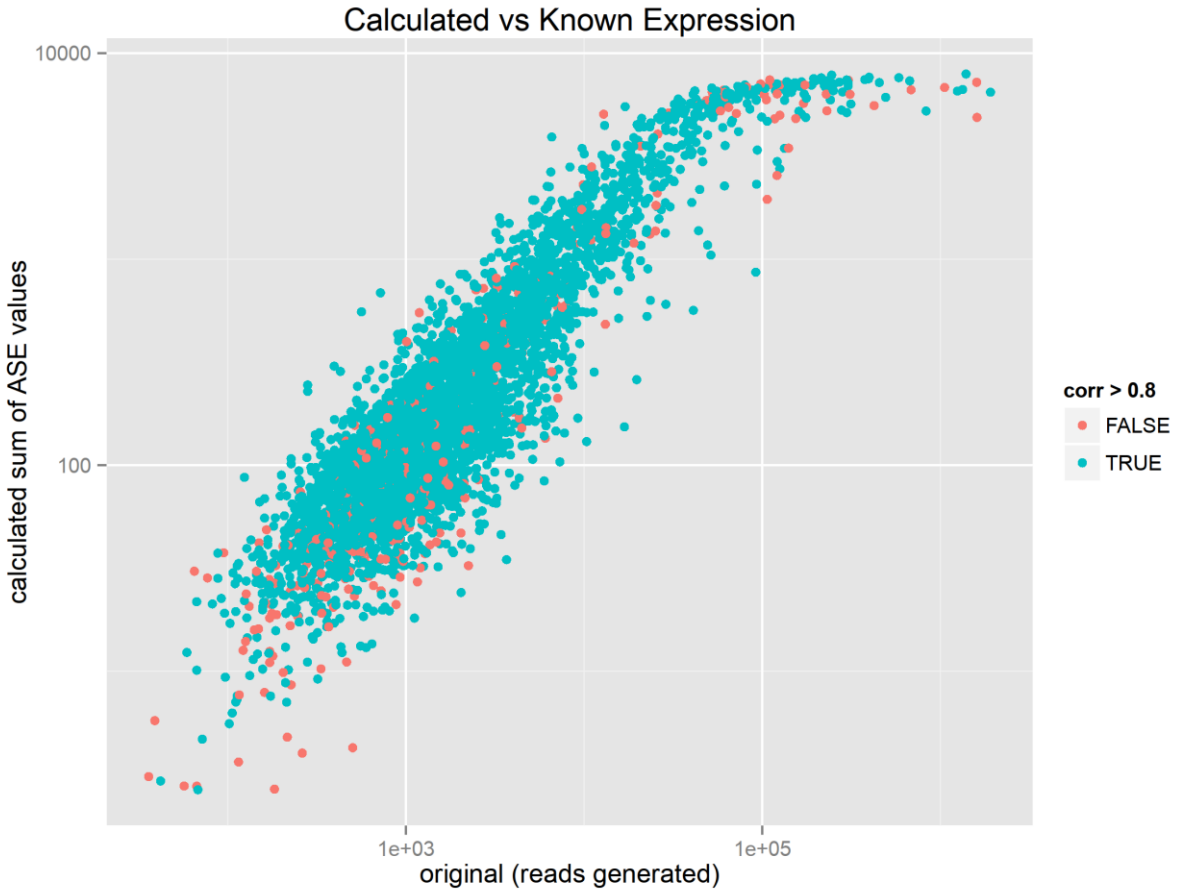



```
# Determine the expression attributable to each allele from the normalized nucleotide counts and expected nucleotides
#   at each dSNP position identified
polyploid_nrmlz_snp_expr.pl --variants= strainA_strainB_strainC.specific.variants --ntcounts=sample.nrml.ntcounts

# Filter out less informative nucleotide positions (0 counts, completely ambiguous, all NTs same, no expression expected)
triploid_snv_filter_Ns.pl --N --nonrep-zero --uncertain --all-same sample.vex > sample.f.vex

# Convert format
updatevex.pl sample.f.vex > sample.f.vex2

# Combine data from useful variant sites into per-gene allele specific expression values.
calc_asge.pl --vex2=sample.f.vex2
```

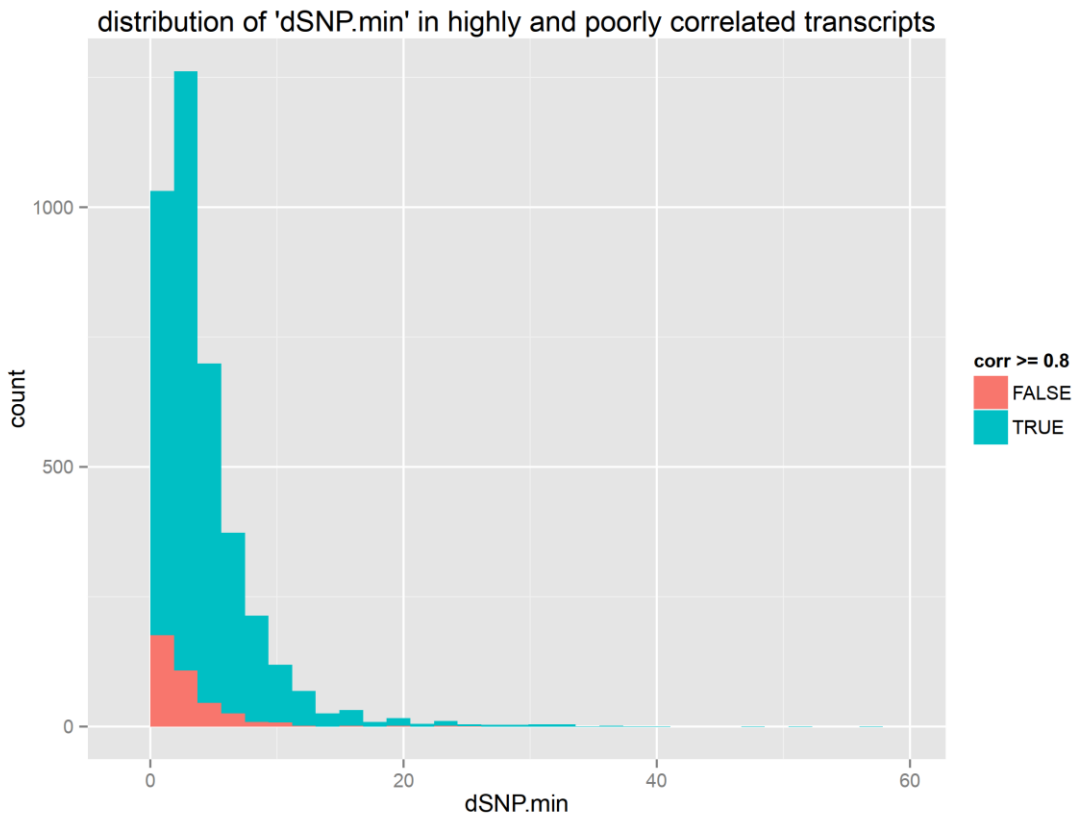
Text S1. Tests to determine association of several factors with low accuracy of ASE. A short summary of extra tests done to investigate possible causes of low accuracy of ASE measurements.

Tests to determine association of several factors with low accuracy of ASE

A plot of total gene expression in calculated vs original data sets with points colored by correlation value shows no obvious pattern. Low and high correlation points are scattered evenly across the distribution



We next examined the effect of transcripts with low dSNP counts. We determined the count of dSNPs representative of the allele with fewest dSNPs for each transcript and called this value a dSNP.min. Then we split transcripts into two groups: those with a correlation coefficient of less than 0.8, and those with one of greater than or equal to 0.8. We then observed the distribution of min.snp (minimum dSNP count) in each group as a histogram. The distributions are not obviously different.



We also examined whether the *mappability* (subsequences which are highly similar to others in the reference set) played a role. We used the GEM-mappability program from the GEM toolkit (ref) to identify sub-sequences with a higher frequency within the reference set. We searched for 31bp regions and processed the output to identify the highest *mappability* score in any subsequence of a given reference transcript. The table below shows the distribution of peak *mappability* scores in the set of transcripts which had low correlation to the original data. The higher the score the more likely it would be that a read aligner could have trouble aligning to at least part of the target sequence. The majority of these problematic transcripts do not get above a *mappability* score of 3 so this does not seem to be a major factor in general.

	peak *mappability* score									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**47**	**83**
Frequency	165	132	50	17	7	5	2	1	1	1

CHAPTER 4

Gene expression dosage regulation in an Allopolyploid fish

Matos I, Machado MP, Schartl M, Coelho MM. Gene Expression Dosage Regulation in an Allopolyploid Fish. Semsey S, ed. PLoS ONE. 10(3):e0116309. (2015)

PLOS ONE

RESEARCH ARTICLE

Gene Expression Dosage Regulation in an Allopolyploid Fish

I Matos[1,2], M. P. Machado[1,2], M. Schartl[2,3], M. M. Coelho[1]*

1 CE3C—Centre for Ecology, Evolution and Environmental Changes, Faculdade de Ciências, Universidade de Lisboa, 1749–016 Lisboa, Portugal, **2** University of Würzburg, Biozentrum, Physiological Chemistry, Am Hubland, Würzburg, Germany, **3** Comprehensive Cancer Center, University Clinic Würzburg, Josef Schneider Straße 6, 97074 Würzburg, Germany

* mmcoelho@fc.ul.pt

OPEN ACCESS

Citation: Matos I, Machado MP, Schartl M, Coelho MM (2015) Gene Expression Dosage Regulation in an Allopolyploid Fish. PLoS ONE 10(3): e0116309. doi:10.1371/journal.pone.0116309

Academic Editor: Szabolcs Semsey, Niels Bohr Institute, DENMARK

Received: October 28, 2014

Accepted: November 21, 2014

Published: March 19, 2015

Data Availability Statement: Files containing the clean reads for S. alburnoides juvenile transcriptome assembly and the clean reads for the S. alburnoides liver transcriptome mapping are available in ArrayExpress, accession number E-MTAB-3174.

Funding: Funding for this study was provided by Fundação para a Ciência e Tecnologia (http://www.fct.pt/index.phtml.pt): [Project PTDC/BIA-BIC/110277/2009 to MC, PhD grant SFRH/BD/61217/2009 to IM and PhD grant SFRH/BD/73335/2010 to MPM]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

How allopolyploids are able not only to cope but profit from their condition is a question that remains elusive, but is of great importance within the context of successful allopolyploid evolution. One outstanding example of successful allopolyploidy is the endemic Iberian cyprinid *Squalius alburnoides*. Previously, based on the evaluation of a few genes, it was reported that the transcription levels between diploid and triploid *S. alburnoides* were similar. If this phenomenon occurs on a full genomic scale, a wide functional "diploidization" could be related to the success of these polyploids. We generated RNA-seq data from whole juvenile fish and from adult livers, to perform the first comparative quantitative transcriptomic analysis between diploid and triploid individuals of a vertebrate allopolyploid. Together with an assay to estimate relative expression per cell, it was possible to infer the relative sizes of transcriptomes. This showed that diploid and triploid *S. alburnoides* hybrids have similar liver transcriptome sizes. This in turn made it valid to directly compare the *S. alburnoides* RNA-seq transcript data sets and obtain a profile of dosage responses across the *S. alburnoides* transcriptome. We found that 64% of transcripts in juveniles' samples and 44% in liver samples differed less than twofold between diploid and triploid hybrids (similar expression). Yet, respectively 29% and 15% of transcripts presented accurate dosage compensation (PAA/PA expression ratio of 1 instead of 1.5). Therefore, an exact functional diploidization of the triploid genome does not occur, but a significant down regulation of gene expression in triploids was observed. However, for those genes with similar expression levels between diploids and triploids, expression is not globally strictly proportional to gene dosage nor is it set to a perfect diploid level. This quantitative expression flexibility may be a strong contributor to overcome the genomic shock, and be an immediate evolutionary advantage of allopolyploids.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

In polyploid lineages resulting from hybridization (allopolyploids), the combination of homeologous chromosomes from divergent species promotes a multitude of biological events [1]. Heterozygosity, divergence of duplicate genes, and novel gene interactions lead to genetic and phenotypic variability [2] that are stably and successfully maintained in these lineages [1]. Allopolyploids are, in this scope, great evolutionary projects full of opportunities for selection and adaptation. On the other hand, allopolyploid lineages have to face an important challenge, namely to overcome genomic shock caused by the simultaneous high level of heterozygosity (due to hybridization) and gene dosage increase (due to polyploidy) [3]. However, mostly plants and invertebrates but also lower vertebrates, deal with these challenges very successfully [4] as they survive and perpetuate. The evolutionary success of several animal allopolyploid lineages like *Squalius alburnoides* [5], *Rana esculenta* [6], *Bufo viridis* [7] or *Poecilia formosa* [8], outdates research that suggests that the fate of (allo)polyploids is a rapid extinction, and suggests that such animals might developed mechanisms that stabilize their genomes as already widely reported in plants [9].

In allopolyploid plants, the reduction of gene redundancy towards a functional diploidization (dosage compensation) has been pointed out as a way to cope with gene dosage increase [10], but in vertebrates this hypothesis has been scarcely investigated. However, the recent recognition that hybridization and polyploidy are much more frequent in animals than previously inferred [11] and that this might have significantly shaped vertebrate genomes [4] highlighted the importance to extend these studies further than to allopolyploid plants and invertebrates. In a first attempt to study gene expression regulation in a vertebrate allopolyploid context, the expression level of 7 genes (gene set encompassing tissue specific and housekeeping genes), were evaluated and the occurrence of a compensation mechanism was reported in the allopolyploid cyprinid *Squalius alburnoides* [12]. In this fish, for those first analysed genes, the existence of a dosage compensation mechanism that brings transcript levels in triploids to the diploid state was shown. Yet, the genomic extension of the phenomenon remains unknown. It may be a global gene dosage compensation event, acting without exception throughout the whole genome. On the other hand, a certain number of genes may escape the dosage-regulation mechanism, or dosage compensation may be restricted to specific subsets of genes, or any other still unformulated conjecture [12]. However, either taking place on a full genomic scale or only partially, the occurrence of similar transcription levels between diploids and triploids can be a relevant factor contributing to the success and perpetuation of polyploids among lower vertebrates. Analyses of entire transcriptomes (RNA-seq), available in the meantime, are now the imminent choices to disentangle this kind of questions, [13, 14, 15]. RNA-seq allows in a fast and cost-effective way to do a simultaneous qualitative and quantitative analysis of complex transcriptomes [16]. It showed to be a general improvement compared to microarrays [17] and was extensively validated by qPCR, exceeding it in range [14, 18].

Squalius alburnoides is an allopolyploid cyprinid, resulting from interspecific hybridization between females of *Squalius pyrenaicus* (P genome) and males of a now extinct species related to *Anaecypris hispanica* (A genome) [5, 19] (Fig. 1A). *S. alburnoides* natural populations are composed of animals of different ploidy levels and genomic constitutions (Fig. 1B) referred to as genomotypes, and include fertile sexual and non-sexual forms (Fig. 1C) Currently, in the Iberian southern basins the predominant *S. alburnoides* genomotypes are the hybrid triploid PAA, diploid PA, and the parental-like diploid AA. Individuals of AA genomotype are all males, and are only reconstituted from hybrids within the complex [5], so despite the homogenomic *A. hispanica*-like nucleus, they carry P mitochondrial DNA and are called nuclear non-hybrids (Fig. 1C).

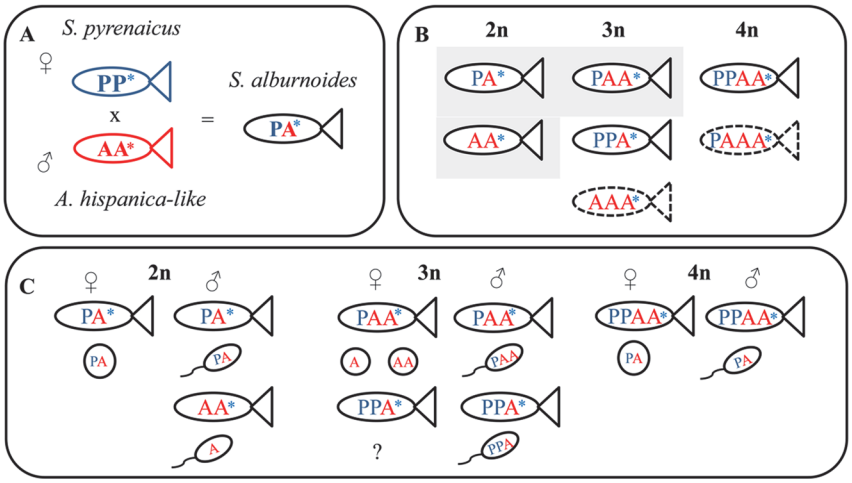


Fig 1. *S. alburnoides* complex simplified overview. A) Initial hybridization event in the origin of the complex. B) Diversity of genomotypes found in the main southern portuguese river basins. In gray background are the *S. alburnoides* genomotypes more frequent in nature, which are in focus in this work. Dashed lines indicate naturally occurring but rare genomotypes. C) Diversity of gametes, produced through a variety of mechanisms—clonally, by hybridogenesis, meiotic hybridogenesis or normal meiosis, depending on the sex and genomic composition of the individual. Asterisk represents mitochondrial genotype: blue from *S. pyrenaicus* and red from *A. hispanica*-like.

doi:10.1371/journal.pone.0116309.g001

In this work, our goal is to expose how an evolutionary successful allopolyploid vertebrate, the cyprinid *S. alburnoides*, deals on the transcriptional level with the genomic stress derived from hybridization and polyploidy. Also, this study aims to contribute to understand the role of gene dosage compensation in the *S. alburnoides* breeding complex.

In the present work, the quantitative expression profiles of diploid and triploid *S. alburnoides* were compared after RNA-sequencing and *de novo* assembly of the *S. alburnoides* transcriptome. Since it is known that there are many transcriptional changes from juveniles to adults [20] and that gene expression patterns can be tissue-specific [12, 21], RNA-seq was done both from whole bodies and at a single tissue level.

However, the RNA-seq transcript profiling experiments the differences in expression of a gene between two samples are in fact differences in expression per unit of RNA or "per transcriptome. To directly infer global expression dosage responses from the RNA-Seq transcript profiling experiments the transcriptomes compared must be of equal size. Without information about the sizes of the transcriptomes compared, direct assumptions about the expression per gene copy or expression per cell drawn from the transcriptome-normalized expression can be flawed [22; 23]. Based only on the expression per transcriptome, the differences in expression per cell that are proportional to the change in the total transcriptome size will appear as equal expression per transcriptome [23]. Consequently, we also estimated the relative transcriptome size from liver samples between diploid and triploid *S. alburnoides* hybrids.

Materials and Methods

Fish samples and genotyping

Adult specimens. From the area of sympatry of *S. pyrenaicus* and *S. alburnoides*, in southern Portuguese river basins, a total 20 specimens (6 *S. pyrenaicus* and 14 *S. alburnoides*) were collected to perform experimental crosses and provide adult biological material. Individuals were collected from tree locations: Algarve basin, Almargem stream (29 S; 622495.24 m E;

4113964.49 m N (UTM)); Guadiana basin, Oeiras stream (29 S; 604985.29 m E; 4164883.94 m N (UTM)) and Tejo basin, Cobrão stream (29 S; 606212.42 m E; 4398531.98 m N (UTM)). Sampling locations were chosen according with the legal permits and considering the existence of a differential geographical distribution of genomotypes, different relative frequencies of each genomotype in each river basin and the fact that diploid and triploid *S. alburnoides* are not morphologically distinguishable [5]. All specimens were adults, sexually mature (determined by obvious abdominal distension and gametes releasing upon slight abdominal pressure) and approximately one to three years old (estimated from the length of each specimen). Fish were captured by electrofishing and brought alive to the laboratory. Each fish was photographed to posterior identification by Scaleprinting [24]. Also, DNA was obtained from fin clips and the specimens were genotyped according to [25]. Fish were acclimated for three weeks in high-quality glass tanks (30 l capacity) equipped with filtration units, under the same standard conditions of light (14 hours light, 10 hours dark), temperature (22°C ±1 °C), water quality (pH between 6.5 and 7.30) and nutrition (twice a day feeding with frozen anthemia and commercial fish food flacks).

Fourteen individuals (6 PA; 6 PAA; 1 AA and 1 PP) were sacrificed and organs were dissected and preserved in RNA later (Ambion) at -20°C.

Juveniles. During the reproductive season, and before its use in the previous section, all adults of *S. alburnoides* and *S. pyrenaicus* (previously genotyped) and visibly sexually maturated were used to perform defined experimental crosses in order to obtain progenies specifically with PAA, PA, AA and PP genomotypes [5, 26]. For each cross, eggs and sperm were collected (by gentle abdominal pressure) from the selected individuals and used for embryo production in petri dishes. Successful fertilization was assessed by observation under the stereoscope after 2h. Viable progeny wit PP genomotype was not obtained. At least 2 viable progenies of each, putatively of AA, PA and PAA genomotypes were reared in high-quality glass tanks (5 l capacity) equipped aerating units, under the same standard conditions of light (14 hours light, 10 hours dark), temperature (20°C ±1°C), water quality (pH between 6.5 and 7.30) and nutrition (twice a day feeding with commercial powder food for fish larvae). 30 days after hatching (dah) several siblings from the same clutch and from each genotype were collected and preserved in RNA later (Ambion) at-20°C. Simultaneous extractions of DNA and total RNA were performed with the AllPrep DNA/RNA Mini Kit (Qiagen). The extracted DNA was used to assess the genomic composition of the selected progenies according to [25].

Library construction and sequencing

Total RNA extracted with the AllPrep DNA/RNA Mini Kit (Qiagen) was DNase treated on-column with the RNase-free DNase Set (Qiagen). At least 15 μg of RNA were obtained per sample. Integrity evaluation and quantification of the extracted RNA was performed with Nanodrop 1000 (Thermo Scientific) and 2100 Bioanalyser (Agilent Technologies). All samples presented a RIN higher than 8.5 (Bioanalyser). Normalized juvenile cDNA libraries were homemade prepared according to [27]. Non-normalized liver libraries were prepared with TruSeq RNA Sample Preparation Kit (Illumina) according to the Illumina specifications. All libraries were paired-end sequenced using Illumina HiSeq 2000.

Juvenile samples. Three barcoded RNA libraries were constructed: juvenile-AA, juvenile-PA and juvenile-PAA. For the construction of the 3 libraries RNA was purified from whole bodies of pools of 4 larvae of each genomotype at 30 dah. At this age, all major organs (except the reproductive systems that are not yet fully defined) are already formed in all 3 investigated genomic forms of *S. alburnoides* (unpublished data). We did pooling of individuals in order to obtain the minimal amount of RNA required for library construction and sequencing. For each

library only siblings from the same cross were used. The 3 libraries were sequenced producing 12 Gb clean data (~ 4Gb per library) in 3 data sets (juv-AA; juv-PA; and juv-PAA) of Illumina HiSeq short paired-end sequence reads (90 bp). The output statistics of sequenced data is available as S1 Table.

Liver samples. Four barcoded RNA libraries were constructed: one for *S. pyrenaicus* (liver- PP) and three for *S. alburnoides* (liver-AA, liver-PA and liver-PAA). For the construction of the libraries RNA was purified from livers, independently for each sample/library. The 4 libraries were sequenced producing 4Gb of clean data (~1Gb per library) in 4 data sets (liv-AA; liv-PA; liv-PP and liv-PAA) (short paired-end sequence reads around 50 bp). The output statistics of sequenced data is available as S2 Table.

Processing of RNA-seq data for gene expression quantification

The raw data of juv-AA, juv-PA, and juv-PAA were processed into clean data by removing reads with adaptors, reads with more than 5% of unknown nucleotides and reads where more than half of the bases' quality values were less than 5. Also, orphan reads were excluded.

Transcriptome de novo assembly was carried out with Trinity [28]. Assemblies were taken into further processes of sequence splicing and redundancy removing with the sequence clustering software TGICL [29]. After clustering, UniGenes were divided in two classes: clusters (prefix CL) and singletons (prefix unigene) (Statistics of assembly quality provided as S3 Table). blastx alignment (e-value < 0.00001) between unigenes and protein databases (nr, Swiss-Prot, KEGG, COG) was performed, and the best aligning results were used to decide sequence direction. When results of different databases conflicted, the priority order of nr, Swiss-Prot, KEGG and COG was followed (statistics of annotation results provided as S4 Table). Unigenes that were not aligned to any of these databases were scanned by ESTScan (v2.1) [30], to decide the sequence direction. The expression level of each unigene was calculated as FPKM, defined as fragments per kilobase of exon model per million mapped fragments [31], with the Cufflinks package (v0.9.3).

Concerning raw sequencing data of liv-AA, liv-PA, liv-PP and liv-PAA, quality filtering was performed: low quality (phred score<20), and ambiguous nucleotides were trimmed off and the quality assessed using FastQC v0.10.1. Reads were mapped to the *Danio rerio* reference genome (Ensembl *Danio rerio* genome Zv9.69) using Stampy v1.0.21 (substitution rate of 11% and no multiple hits allowed) (Mapping statistics are presented as S2 Table). To use the mapping approach the divergence rate between *D. rerio* and *S. pyrenaicus* (11%) and *D. rerio* and *S. alburnoides* AA genomotype (10%) had to be assessed and the higher value was used (11%). Fragments mapped into genes were counted using HTSeq v0.5.3p9 (htseq-count option for no stranded data). FPKM values were calculated using fragment counts from HTSeq and total fragments mapped obtained with Samtools v0.1.18. (flagstat option, counting pair reads plus singletons mapped). Differentially expressed genes were calculated using Bioconductor edgeR package v3.0.8 (for a FDR<0.05) after data normalization using Bioconductor EDASeq package v1.4.0 (first normalized to gene length and second to the libraries size).

Comparative analysis of expression levels

Expression differences were obtained by dividing the normalized expression values (FPKM) in one library by the corresponding expression value of the same transcript in each other library (fold change) independently for juveniles and livers. The quantitative comparative profiles were displayed through orderly plotting of $\log_2$ (fold change). The value of $|\log_2(\text{Ratio})|<1$ was considered to be the threshold for similar expression. A false discovery rate (FDR) ≤ 0.05 was

used as cutoff threshold to determine the significance of differential expression (FDR correction, version for dependent tests, applied to the raw p-value of all transcripts).

Also, the observed gene expression level of each transcript in the hybrids was compared to an expected expression level if P and A alleles are expressed exactly at the same level as in the non-hybrid situation (additivity expectation). The expression level of each gene in the parental diploids (AA and PP) was used to calculate the expected additive expression for each gene in both hybrids. Then, the observed expression value (obs) of each transcript was divided by its corresponding expected additive value (exp), both in liv-PAA and in liv-PA. These ratios were $\log_2$ transformed and when within the interval $-1<\log_2(\text{observed FPKM/expected FPKM})<1$, the transcripts were considered as additively expressed.

When appropriate, the χ-square test was applied to the data and is indicated.

Functional analysis

Functional enrichment analyses were carried out using DAVID Bioinformatics Resource 6.7 (http://david.abcc.ncifcrf.gov/). The top blastx hits in nr database corresponding to each *S. alburnoides* unigene were used as customized reference background for juveniles' data set. As liver gene expression reference background we used the mapped genes with expression in at least one of the 4 liver libraries. DAVID sorting thresholds were changed to EASE score (modified Fisher exact test) ≤ 0.01. Significant enrichment was only considered when Benjamini corrected p-value ≤ 0.05.

Data accessibility

Files containing the clean reads for *S. alburnoides* juvenile transcriptome assembly and the clean reads for the *S. alburnoides* liver transcriptome mapping are available in ArrayExpress, accession number E-MTAB-3174.

qRT-PCR genome-normalized expression assay and relative transcriptome size estimation

RNA and gDNA (total nucleic acid [TNA]) were co-extracted from RNA later (Ambion) preserved livers according to the extraction protocol described in [32] with minor modifications. TNA were extracted independently from 5 livers of each diploid and triploid hybrid (PA and PAA; total n = 10) and the presence of both RNA and gDNA confirmed in a 1,5% agarose gel for all 10 samples. 1 μg of TNA per sample was reversed transcribed with RevertAid First Strand cDNA Synthesis Kit (Fermentas) with oligo dT primers.

Primers for target genes were designed to be specific to either cDNA or gDNA (S5 Table). For cDNA-specific primers, one or both primers in a pair were designed to span exon-exon splice junctions and for gDNA-specific primers, one or both primers were designed to prime at least partially within an intron. Template specificity was confirmed for all primer pairs by qPCR with cDNA and gDNA templates. Primers specific to cDNA were designed for 6 genes (*rpl8*, *rpl35*, *actb2*, *pabpc1a*, *eef1a* and *rpsa*) and primers specific to gDNA were designed for 3 genes (*rpl8*, *eef1a* and *actb2*) (S5 Table).

The cDNA/gDNA mix was diluted 1:1 in nuclease free-water and was 1 μl used as template for each qRT-PCR reaction.

Real-time PCR reactions were performed in BioRad's CFX96 Real-time PCR system (C1000 Thermal Cycler). Real-time PCRs were done in a final volume of 10 μl, with SsoAdvanced Universal SyberGreen Supermix (BioRad) in accordance to the specification of the supplier. The thermal cycling protocol was as follows: initial denaturation step at 95°C for 30s, followed by 40 cycles at 95°C for 10s and 60°C for 30s. For each primer pair, we amplified three technical

replicates from each of the five biological replicates of each of the two genomotypes. Expression of each target gene (cDNA-specific amplification) was normalized to the geometric mean of amplification from the three gDNA-specific targets. Relative genome-normalized expression values were calculated by the Livak method [33].

We calculated expression per cell in PAA relative to PA as 1,5x the relative expression per genome.

RPKM values from the liver data set, obtained as described above, were taken as the transcript abundance per transcriptome for any given gene.

To estimate the hybrid triploid liver transcriptome relative to the hybrid diploid liver transcriptome the per cell expression ratios from the qRT-PCR assay were divided by the per transcriptome expression ratios from the liver RNA-seq data set.

Ethics Statement

Fish captures and handling needed the permission of Instituto de Conservação da Natureza e das Florestas (ICNF), the Portuguese national authority and relevant body concerned with protection of wildlife. ICNF considered that our study was not intrusive issuing the permits AFN, fishing credential n° 82/2012 and ICNB license n°142/2012/CAPT. The studied species are not endangered or protected, nevertheless the selected populations for the captures ware not imperiled, and sampling was done avoiding depletion of the natural stock. Electrofishing was performed in low duration pulses to avoid killing juveniles (300 V, 2–4 A) and the transport to the laboratory was made in appropriate aerated containers. The maintenance and use of animals in the animal facility of the Faculdade de Ciências da Universidade de Lisboa had the approval of the Direção-Geral de Veterinária, Direção de Serviços de Saúde e Proteção Animal (DGV-DSSPA), stated in the "ofício circular" n° 99 0420/000/000/9/11/2009. Fish were handled following the recommended ethical guidelines in [34] and at all times, all efforts were made to minimize suffering. All required manipulations for identification, genotyping and to accomplish the experimental crosses were performed under light anesthesia (40 ppm of MS222 dissolved in the water). Fish used for the experimental crosses that were not sacrificed were later returned to their original capture site. The sacrificed individuals were submitted to an overdose of MS222 (400 ppm of MS222 dissolved in the water) and kicky decapitate previously to the organs harvesting to guarantee the death prior to the harvesting.

Results

Comparative expression profiling from triploid and diploid juveniles

To investigate if the quantitative expression profile of mRNAs changes with ploidy increase we made pairwise comparisons between the expression level profiles of each pair of triploid vs diploid juvenile *S. alburnoides* genomotype. We plotted the $\log_2$ FPKM ratios (juv-PAA/juv-PA) (Fig. 2A) and $\log_2$ FPKM ratios (juv-PAA/juv-AA) (Fig. 2B), producing crescent curves where positive values indicate mRNAs with higher expression in juv-PAA than in juv-PA or juv-AA, and negative values represent mRNAs with lower expression in juv-PAA then in juv- PA or juv-AA respectively. The same comparative expression profiling was performed between juv-PA/juv-AA (Fig. 2C) to illustrate a comparison at the diploid level.

We observed a significant (χ-test, $p<0.001$) higher amount of lower expressed unigenes in juv-PAA compared to juv-PA (Fig. 2A) and also (χ-test, $p<0.001$) in the comparison of juv-PAA with juv-AA (Fig. 2B). This is contrary to what would have been expected from a dosage effect between triploid and diploid organisms. For a dosage effect it would be expected that most transcripts would be higher represented in triploids then in diploids, or similarly represented in case of dosage compensation. Concerning the comparison between juv-PA and juv-

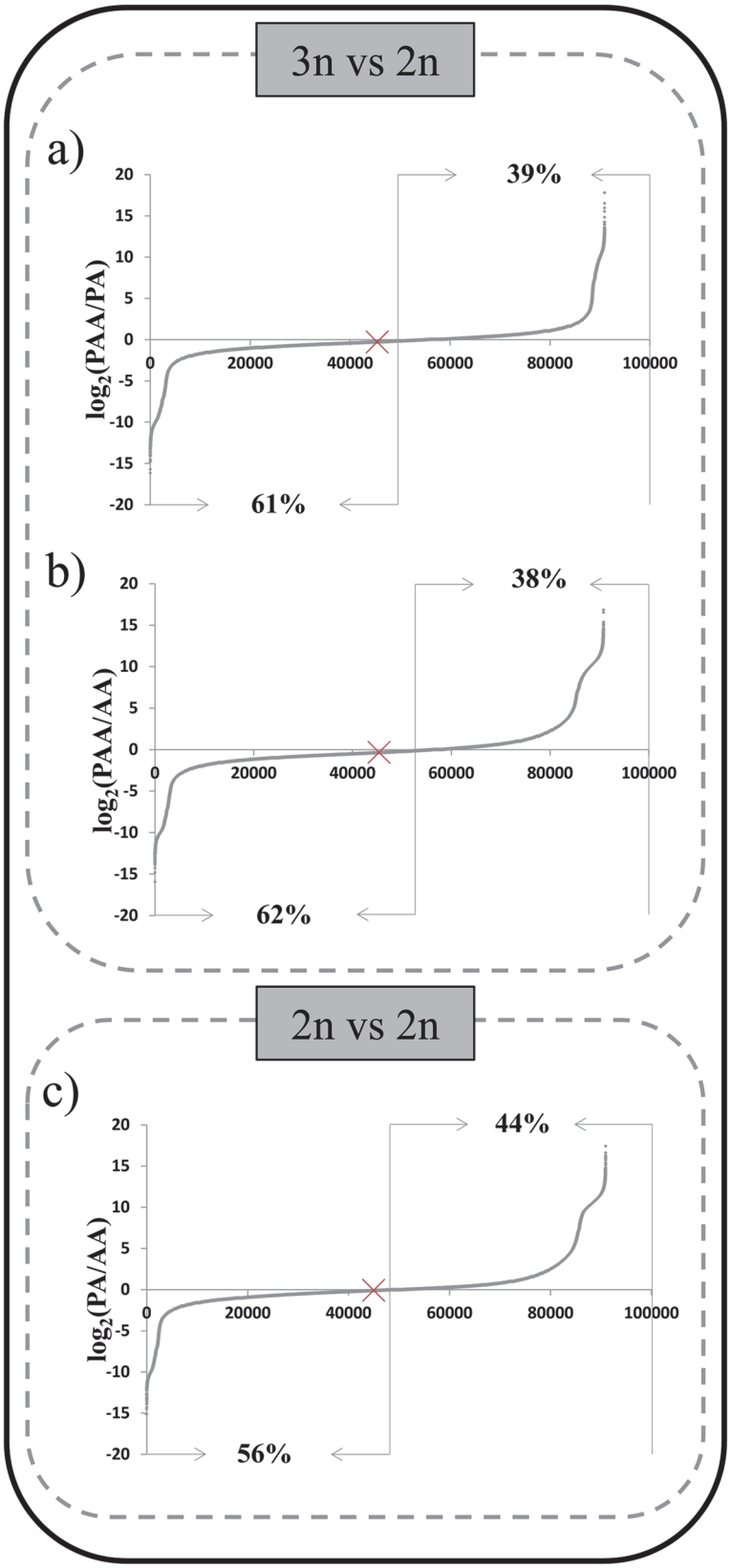

3n vs 2n
a)
39%
61%
log2(PAA/PA)
b)
38%
62%
log2(PAA/AA)
2n vs 2n
c)
44%
56%
log2(PA/AA)
20
15
10
5
0
-5
-10
-15
-20
0
20000
40000
60000
80000
100000

Fig 2. Comparative gene expression profiles between three juvenile genomotypes of the *S. alburnoides* complex. Logarithmized ratios of gene expression for each unigene were orderly plotted producing characteristic crescent curves where positive values indicate transcripts with higher expression and negative values transcripts with lower expression. Median is marked with a cross and indicates if most values are positive or negative. For all comparisons, the difference in the number of lower vs higher expressed transcripts is significant (χ-test, $p<0.001$). The percentages of positive and negative values in each comparison are indicated.

doi:10.1371/journal.pone.0116309.g002

AA, where ploidy rise has no part (only hybridization), despite significant (χ-test, $p<0.001$) a much less conspicuous difference in the number of lower vs higher expressed transcripts is observed (Fig. 2C).

Focusing only on the ploidy level effect, we compared the amount of unigenes that are lower (n = 55545) and higher (n = 35411) expressed between juv-PAA and juv-AA with those lower (n = 50942) and higher (n = 40004) expressed between juv-PA and juv-AA. The expression pattern of the hybrids compared to a diploid non-hybrid is significantly affected (χ-test, $p<0.001$) by the ploidy level of the hybrids. However, the expression profile of the other parental diploid genomic composition (PP) is needed to make a firm conclusion.

Comparative expression profiling of livers from diploid and triploid adults

Unlike for juveniles, material from adult *S. pyrenaicus* (PP) could be easily obtained. Therefore, a second set of quantitative expression data, using adult tissues (livers) was produced.

As for the juveniles, the expression profiles for the liver were obtained per genomotype, and pairwise comparisons were performed (Fig. 3). The comparison between triploid and diploid levels showed in all cases a significant difference in the number of higher and lower represented transcripts (χ-test, $p<0.001$) (Fig. 3A-3C), indicating that globally, ploidy level affects the quantitative expression pattern. Despite the higher gene dosage in the triploid, there is a higher amount of lower represented transcripts in liv-PAA compared to both diploid liv-PA and liv-AA (Fig. 3A-3B). This was consistent with what was observed in the whole body juvenile data set (Fig. 2). Moreover, there is a substantially high amount of higher represented transcripts in liv-PAA when compared to liv-PP (Fig. 3C).

The quantitative relative expression patterns within the same ploidy level (2n) were also inspected (Fig. 3D-3F). First, in the comparison between the two parental diploid genomotypes (AA vs PP), it was observed that a massive amount of transcripts were represented at higher levels in AA liver than in PP liver (Fig. 3D). Then, when comparing PA with each one of the parental genomotypes (AA and PP), we observe that in comparison with liv-AA, the difference between lower and higher represented transcripts is only marginal (Fig. 3E), while in the comparison with liv-PP it is really high (Fig. 3F).

Additivity

We observed that for liv-PAA only 36% of transcripts are represented in the range of the expected/additive expression level, and from the transcripts that are not additively expressed, a very significant (χ-text, $p<0.001$) majority (56% of the total) are under-expressed compared to the additivity expectations (Fig. 4). Hence, in triploid hybrids gene expression in the liver is mostly negatively non-additive.

In the case of the diploid hybrid (PA), we observed that the percentage of additively expressed transcripts rises to more than half of the transcripts (60%) (Fig. 4). From the non-additively expressed ones a significant (χ-text, $p<0.001$) majority (30% of the total) is also under-expressed in comparison to the additivity expectations (Fig. 4).

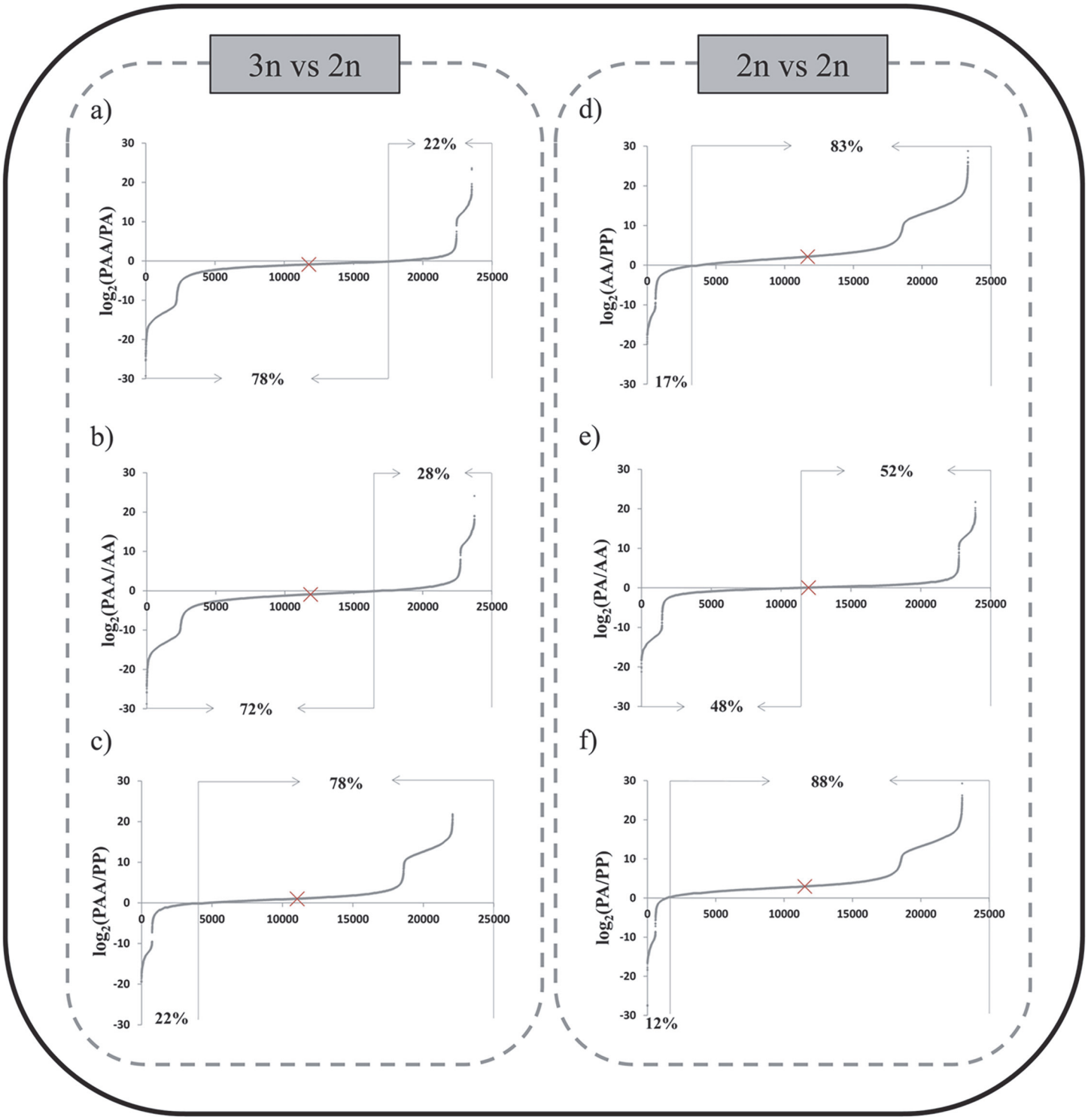


Fig 3. Comparative gene expression profiles in adult liver between the most common forms of the *S. alburnoides* complex. Logarithmized ratios of expression for each mapped transcript ware orderly plotted producing a crescent curve where positive values indicate transcripts with higher expression and negative values transcripts with lower expression. Median is marked with a cross and indicates if most values are positive or negative. For all comparisons, the difference in the number of lower vs higher expressed transcripts is significant (χ-test, $p<0.001$).The percentages of positive and negative values in each comparison are indicated.

doi:10.1371/journal.pone.0116309.g003

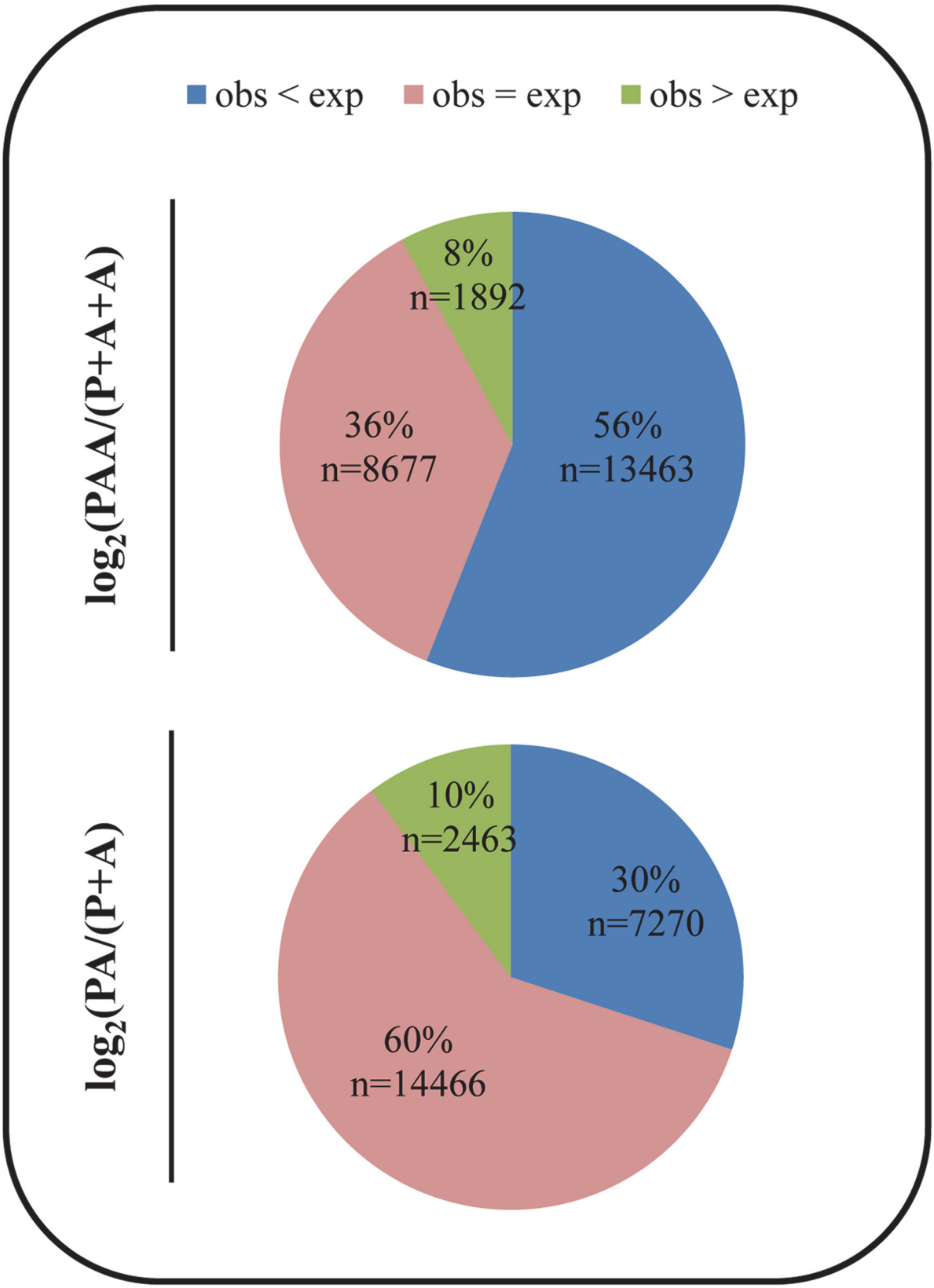

obs < exp
obs = exp
obs > exp
log2(PAA/(P+A+A))
8%
n=1892
36%
n=8677
56%
n=13463
log2(PA/(P+A))
10%
n=2463
30%
n=7270
60%
n=14466

Fig 4. Gene expression additivity in hybrids. Additivity for each transcript was calculated by dividing the observed FPKM value by the expected FPKM value. The expected values were obtained using the expression values of the non-hybrid diploids—AA and PP. The expected value for PAA is (PP/2)+(AA/2)+(AA/2) and the expected value for PA is (PP/2)+(AA/2). Transcript were than evaluated as having lower observed expression then the additivity expectation (obs<exp), observed expression similar to the additivity expectation (obs = exp) or higher observed expression then the additivity expectation (obs>exp). The percentages of transcripts in each category, for both a) PA and b) PAA are represented.

doi:10.1371/journal.pone.0116309.g004

Similar gene expression and dosage compensation

We quantified the similarly expressed transcripts (SE) between each pair of 3n vs 2n genomotypes, both in juveniles and liver data sets (S6 Table). Focusing on the comparisons between hybrids (Table 1), we observe that 64% of the transcripts in juveniles and 44% in the livers are represented similarly in diploids and triploids. Within the SE group we also evaluated the occurrence and/or extension of strictly diploid expression levels in triploids (fold change equal to 1) and also of increased expression in triploids proportional to dosage increase (1.5 fold higher). So, we sorted the SE transcripts into 4 classes. Class I comprises the compensated transcripts, with ratio PAA/PA approximately equal to 1 and within the interval] 0.75;1.25[; in class II are the dosage sensitive ones, with ratio approximately equal to 1.5 and within] 1.25;1.75[; in class III are transcripts affected by some repression, with ratios lower than 0,75 and; in class IV are transcripts overexpressed, with ratios higher than 1.75 (Table 1). The results show that in triploid hybrids more than one third of the SE transcripts are strictly dosage compensated to the diploid hybrid level (45% in juveniles and 35% in liver) (Table 1). On the other hand, there is a much smaller representation of SE transcripts that follow the "1.5-fold rule", being expressed proportionally to gene dosage (17% in juveniles and 13% in liver) (Table 1). Of notice is that in both data sets only a very small percentage of the SE PAA transcripts (4% in juveniles and 3% in livers) present an expression level higher than 1.5 fold (class IV), while 34% in juveniles and 49% in livers are repressed beyond dosage compensation (class III).

Differential expression

Considering the significance criteria for differential gene expression (see material and methods) we quantified the significantly different expressed transcripts between diploid and triploid juveniles and livers (S7 Table). Focusing on the comparisons between hybrids (Table 2), we observed that 22.5% of unigenes in juveniles and 0.83% genes in livers are DE between diploids and triploids. Also, both for juveniles and livers the significant majority of the DE transcripts are higher represented in PAs then in PAAs, despite the higher gene dosage in triploids (Table 2).

Table 1. Similarly expressed transcripts (SE) between triploid and diploid S. alburnoides hybrids in juveniles and livers.

Comparisons	SE (% of total)		SE *per* class	% of SE	% of total
Juveniles *PAA*/*PA*	58076 **(64%)**	**I**	26376	45	**29**
		II	9935	17	**11**
		III	19672	34	22
		IV	2093	4	2
Liver *PAA*/*PA*	10068 **(44%)**	**I**	3508	35	**15**
		II	1308	13	**6**
		III	4947	49	21
		IV	305	3	1

Total numbers and percentages of SE unigenes (juveniles) or mapped genes (livers) and total numbers and percentages of SE's per expression class.

doi:10.1371/journal.pone.0116309.t001

Table 2. Differentially expressed transcripts (DE) between triploid and diploid S. alburnoides hybrids in juveniles and livers.

Comparisons	DE (% of total)	DE group	DE *per* group	% of DE	% of total
Juveniles *PAA/PA*	20468 **(22.5%)**	DEH	6813	33	7,5
		DEL	13655	67	15
Livers *PAA/PA*	195 **(0.83%)**	DEH	41	21	0,17
		DEL	154	79	0. 65

Total numbers and percentages of differently expressed (DE) unigenes in juveniles or mapped genes in livers. DE's were divided in two groups: significantly higher expressed in PAA compared to PA (DEH), and significantly lower expressed in PAA compared to PA (DEL).

doi:10.1371/journal.pone.0116309.t002

Functional enrichment analysis

We used the annotated *de novo* assembled transcriptome of juveniles and the mapping of the *S. alburnoides* liver transcriptome to the Zebrafish genome to perform a functional analysis [35]. In order to look for the biological context for the gene expression dosage regulation observed between diploid and triploid hybrids, we performed a GO and a KEGG pathway enrichment analysis in DE and SE groups.

We found significant functional enrichment in both DE and SE groups in juveniles (Table 3) and livers (Table 4), and the analysis is quit consistent between the two data sets. Briefly, the SE group is enriched in terms associated with metabolic processes, intracellular parts and constitutes of the ribosomes, while DE is mostly enriched in terms associated with the cell membrane (e.g. transport, adhesion, motility). From the KEGG pathway analysis in the SE group we observed that it is consistent with an enrichment of ribosomal components and ribosomal-linked pathways in both data sets. KEGG pathway analysis of the DE group of juveniles is significantly enriched in components of the circadian rhythm, Wnt signaling and melanogenesis pathways. The DE group in the liver data set is significantly enriched in components of the sphingolipid metabolism and PPAR signaling pathways. Within the SE group we also looked for differential functional enrichment between classes I and II in both juveniles and liver data sets (Table 5). Within our criteria of significance, no significant functional enrichment was detected in class I, both in juveniles and livers. Class II of juveniles is enriched in terms linked to ribosomal complex, to the respiratory chain and to the hemoglobin complex. Livers class II is enriched in terms linked to ribosomes.

Relative transcriptome size

To estimate the relative size of the PAA transcriptome vs the PA transcriptome we used livers from both hybrid genomotypes and analyzed six target genes (*rpl8*, *rpl35*, *actb2*, *pabpc1a*, *eef1a* and *rpsa*) to obtain genome-normalized expression estimates through a qRT-PCR assay and transcriptome-normalized expression estimates from the RNA-Seq assay.

In order to estimate relative expression level per genome, we used a qRT-PCR strategy devised by [23] that normalizes cDNA amplification to genomic DNA (gDNA) amplification. The simultaneous RNA and gDNA extraction from the same cells preserves the in vivo RNA/gDNA ratios. This allowed us to normalize gene expression (cDNA amplification) to genome copy number (gDNA amplification), which directly gives the transcript abundance per genome. With this approach we quantified the expression per genome in the allotriploid (PAA) *S. alburnoides* relatively to its diploid counterpart (PA) for the six target genes (Table 6).

Because PAA has three copies of each gene, for every two copies in PA diploids we calculated expression per cell in PAA relative to PA as 1,5x the relative expression per genome (Table 6).

Table 3. Functional enrichment in GO terms and KEGG pathways of PAA vs PA similarly expressed and differentially expressed gene groups for juveniles.

		Term	#	FE	p-val.
SE	BP	catabolic process	326	1,1	4,6E-3
	BP	macromolecule metabolic process	1993	1,0	3,6E-2
	BP	primary metabolic process	2496	1,0	4,7E-2
	CC	intracellular part	2829	1,0	6,7E-4
	BP	ribonucleoprotein complex	195	1,1	7,1E-4
	BP	cell part	4913	1,0	6,9E-4
	BP	intracellular	3450	1,0	6,0E-4
	BP	intracellular organelle	2358	1,0	6,3E-4
	MF	structural constituent of ribosome	94	1,1	3,0E-2
	KEGG	Spliceosome	93	1,2	1,4E-3
DE	BP	**cell adhesion**	85	1,7	2,1E-6
	BP	cellular developmental process	154	1,3	1,5E-2
	BP	**cellular component morphogenesis**	61	1,5	1,0E-2
	BP	anatomical structure development	282	1,2	8,5E-3
	BP	**cell motion**	56	1,5	4,8E-2
	BP	anatomical structure morphogenesis	169	1,2	5,0E-2
	CC	**extracellular region part**	57	1,5	1,2E-2
	CC	**extracellular matrix**	37	1,6	4,7E-2
	MF	signal transducer activity	272	1,3	3,9E-5
	MF	ion binding	620	1,1	1,2E-3
	KEGG	**Circadian rhythm**	12	3,1	3,2E-2
	KEGG	**Wnt signaling pathway**	43	1,5	4,7E-2
	KEGG	**Melanogenesis**	31	1,6	3,9E-2

(BP) Biological process, (MF) molecular function, (CC) cellular component, (KE) KEGG pathway, (#) number of transcripts, (FE) fold enrichment, (p) Benjamini corrected p-value. GO enrichment analysis was performed at the secondary classification of terms. Terms with FE $\geq$ 1.5 [35] are in bold.

doi:10.1371/journal.pone.0116309.t003

The transcript abundance per transcriptome (RPKM values) for all six target genes were searched within the liver data set and the relative expression per transcriptome between liv-PAA and liv-PA was calculated (Table 6, Fig. 5). With this approach we obtained six independent estimates of the size of the triploid transcriptome relative to the diploid hybrid transcriptome. As expected, there was variation among individual gene estimates, but on average the PAA transcriptome was equal in size to the PA transcriptome. With these data we rejected the null hypothesis that the triploid hybrid transcriptome was subjected to a genome-wide dosage effect, as it was not increased 1.5 fold relative to the diploid hybrid transcriptome (P<0,0001; One sample *t*-test). On the other hand, the null hypothesis that PAA transcriptome was equal in size (genome wide dosage compensation) to the PA transcriptome was not rejected (P = 0.8867; One sample *t*-test).

Discussion

Transcriptome size and overall expression

To directly infer global expression dosage responses from the RNA-Seq transcript profiling experiments the transcriptomes compared must be of equal size [23]. When comparisons are made between different ploidy levels, intuitively this assumption is flawed due to the real

Table 4. Functional enrichment in GO terms and KEGG pathways of PAA vs PA similarly expressed and differentially expressed gene groups for livers.

		Term	#	FE	p-val.
SE	BP	cellular metabolic process	1442	1,1	1,0E-16
	BP	macromolecule metabolic process	1259	1,2	5,3E-16
	BP	primary metabolic process	1529	1,1	5,9E-11
	BP	**ribonucleoprotein complex biogenesis**	42	1,8	1,8E-5
	BP	nitrogen compound metabolic process	663	1,1	2,0E-5
	BP	biosynthetic process	656	1,1	1,4E-4
	BP	catabolic process	205	1,2	9,0E-3
	BP	macromolecular complex subunit org.	82	1,3	9,4E-3
	BP	**establishment of RNA localization**	18	1,9	1,9E-2
	BP	**translational initiation**	24	1,7	2,0E-2
	CC	**ribonucleoprotein complex**	185	1,6	2,3E-18
	CC	intracellular	2021	1,1	5,3E-15
	CC	intracellular part	1652	1,1	4,5E-12
	CC	intracellular organelle	1369	1,1	4,7E-9
	CC	**organelle lumen**	138	1,5	1,3E-8
	CC	membrane-bounded organelle	1164	1,1	8,1E-8
	CC	cell part	2750	1,0	2,9E-4
	CC	intracellular organelle part	446	1,1	2,5E-3
	CC	organelle part	446	1,1	2,5E-3
	CC	non-membrane-bounded organelle	322	1,1	4,6E-3
	MF	nucleic acid binding	1002	1,2	1,2E-15
	MF	**structural constituent of ribosome**	104	1,7	9,3E-13
	MF	nucleotide binding	793	1,1	1,5E-7
	MF	**translation factor activity**	60	1,7	1,3E-6
	MF	nucleoside binding	501	1,1	1,6E-3
	MF	transferase activity	630	1,1	1,9E-3
	MF	ion binding	1052	1,1	4,2E-3
	MF	ligase activity	116	1,2	2,0E-2
	KEGG	**Ribosome**	79	2,0	2,1E-19
	KEGG	**Spliceosome**	89	1,7	6,5E-10
	KEGG	Ubiquitin mediated proteolysis	76	1,4	3,6E-3
	KEGG	**RNA degradation**	38	1,5	2,2E-3
DE	BP	**transport**	28	2,7	1,5E-5
	BP	**establishment of localization**	28	2,7	9,1E-6
	BP	**transmembrane transport**	11	3,6	1,1E-2
	CC	**membrane**	34	1,5	3,6E-2
	CC	**apical part of cell**	3	37,5	2,3E-2
	MF	**substrate-specific transporter activity**	17	4,3	1,7E-5
	MF	**hydrolase activity**	26	2,5	1,1E-4
	MF	**transmembrane transporter activity**	13	3,5	1,6E-3
	KEGG	**Sphingolipid metabolism**	4	16,8	2,2E-2
	KEGG	**PPAR signaling pathway**	4	13,3	2,2E-2

(BP) Biological process, (MF) molecular function, (CC) cellular component, (KE) KEGG pathway, (#) number of transcripts, (FE) fold enrichment, (p) Benjamini corrected p-value. GO enrichment analysis was performed at the secondary classification of terms. Terms with FE $\geq$ 1.5 [35] are in bold.

doi:10.1371/journal.pone.0116309.t004

Table 5. Differential functional enrichment in GO terms and KEGG pathways between class I and II, both in juvenile and liver data sets.

		Juveniles			
		Term	**#**	**FE**	**p-val.**
Class I		No significant enrichment			
		Term	**#**	**FE**	**p-val.**
Class II	BP	heterocycle biosynthetic process	21	2,9	3,9E-3
	BP	tetrapyrrole metabolic process	12	4,0	1,2E-2
	BP	tetrapyrrole biosynthetic process	11	4,2	1,2E-2
	MF	structural molecule activity	65	1,6	3,4E-2
	MF	heme-copper terminal oxidase activity	10	4,1	3,7E-2
	MF	cytochrome-c oxidase activity	10	4,1	3,7E-2
	MF	oxidoreductase activity, acting on heme group	10	4,1	3,7E-2
	KEGG	Ribosome	21	2,6	5,7E-3
		Liver			
		Term	**#**	**FE**	**p-val.**
Class I		No significant enrichment			
		Term	**#**	**FE**	**p-val.**
Class II	BP	translation	71	3,0	1,4E-16
	CC	ribosome	62	3,4	3,7E-20
	CC	ribonucleoprotein complex	65	2,3	1,6E-10
	CC	intracellular non-membrane-bounded organelle	82	1,7	6,6E-6
	CC	non-membrane-bounded organelle	82	1,7	6,6E-6
	CC	ribosomal subunit	11	3,8	6,3E-3
	CC	small ribosomal subunit	7	4,6	4,8E-2
	MF	structural constituent of ribosome	60	4,3	8,5E-24
	MF	structural molecule activity	72	3,4	2,9E-21
	KEGG	Ribosome	60	5,4	5,2E-35

(BP) Biological process, (MF) molecular function, (CC) cellular component, (KE) KEGG pathway, (#) number of transcripts, (FE) fold enrichment, (p) Benjamini corrected p-value. GO enrichment analysis was performed at the slim classification of terms.

doi:10.1371/journal.pone.0116309.t005

genome-wide differences in gene dosage. However, in the *S. alburnoides* case, the hypothesis put forward in [12] and explored in the present work, is that there is a common "diploid" state of genic activity between diploid and triploid *S. alburnoides* individuals. Following the method described and implemented in [23], that couples transcript profiling data with a genome normalized qRT-PCR assay we estimated the liver transcriptome size of the *S. alburnoides* triploid hybrid (PAA) relatively to the liver transcriptome size of the diploid *S. alburnoides* hybrid (PA). We showed that the two compared transcriptomes are fairly the same size. This validates the direct use of the RNA-Seq transcript profiling experiments to infer the "gene by gene" global pattern of expression dosage responses between diploid and triploid *S. alburnoides*. Moreover, it supports at an overall scale the previous conjecture of transcriptional equivalence between diploid and triploid *S. alburnoides*.

Allopolyploid genome regulation

One of the puzzling features of *S. alburnoides* complex is the extraordinary morphological similarity between PA and PAA individuals, which are even undistinguishable by morphometric characters [36]. Conversely, PP and AA genomotypes are easily distinguishable from each

Table 6. Data and calculations used for estimating relative triploid vs diploid hybrid transcriptome size.

Gene	[1]Transcripts/ genome (qRT-PCR; N = 5)		[2]Transcripts/cell (qRT-PCR; N = 5)	[3]Transcripts/transcriptome (RPKM; N = 1)	[4]Transcriptome size
	PAA/PA	SD	PAA/PA	PAA/PA	PAA/PA
rpl8	0,8	0,1	1,1	1,5	0,8
eef1a	0,8	0,1	1,1	1,3	0,9
actb2	0,6	0,2	0,9	0,6	1,6
rpsa	0,9	0,3	1,3	1,5	0,9
pabpc1a	0,8	0,2	1,2	1,0	1,2
rpl35	0,8	0,2	1,2	1,5	0,8
media	**0,8**		**1,2**	**1,0**	**1,0**
SD	0,1		0,1	0,3	0,3

[1]-Expression quantified by qRT-PCR, using total nucleic acid as the template for reverse transcription and normalization to genome copy number.

[2]- Because PAA has 3 genomes per cell, meaning 1,5x the amount of genomes per diploid cell, the transcripts/cell values for PAA/PA are equal to 1,5x the values for transcripts/genome.

[3]-For each target gene RPKM values were derived from the liver RNA-Seq data set.

[4]-Trascriptome size is determined by dividing "transcripts/cell" by "transcripts/transcriptome".

doi:10.1371/journal.pone.0116309.t006

other and from the hybrids. The stable phenotypic similarity between PA and PAA hybrids could be interpreted as indication of similar gene expression between them [12, 37]. At its quantitative component, this hypothesis was corroborated on a small scale [12, 38], when diploid and triploid individuals were found to have similar expression levels for a small analyzed gene set. Conversely, our genome wide approach shows that, despite many genes do not present a significant differential expression between diploid and triploid hybrids, PAA gene

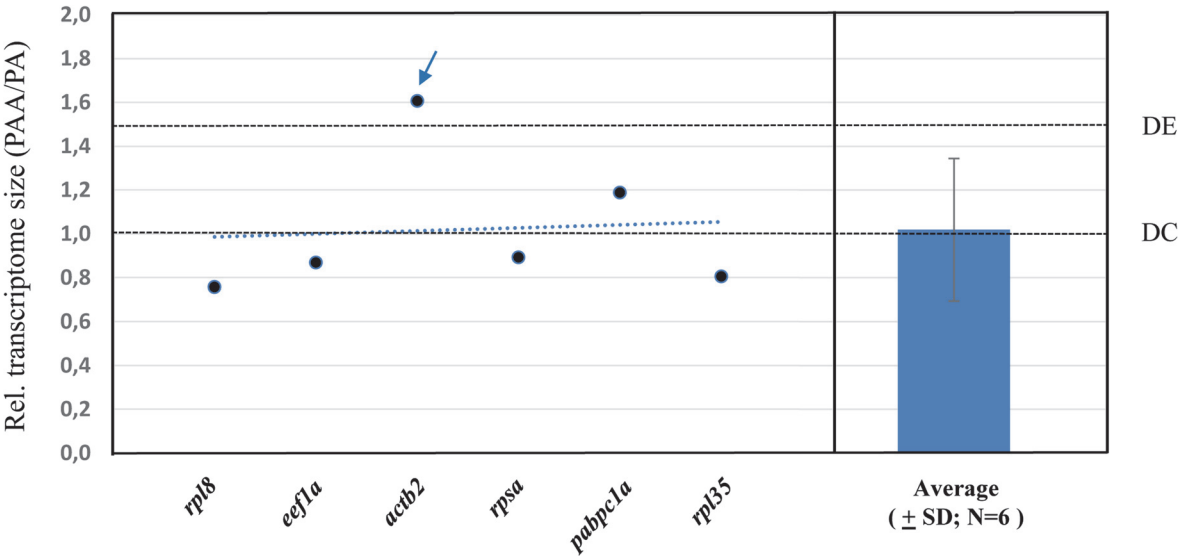


Fig 5. *S. alburnoides* PAA transcriptome size relative to the transcriptome of the diploid PA. Six individual gene-based estimates of relative transcriptome size and average estimate (±SD; N = 6) of the triploid hybrid transcriptome relative to the diploid hybrid transcriptome. DE represents the expected value if the PAA transcriptome experienced a genome-wide dosage effect. DC represents the expected value if PAA transcriptome experienced genome-wide dosage compensation. Dashed blue line represents the tendency curve of the sactterplot. The blue arrow indicates a possible outlier in the data set but that hypothesis was rejected (Tukey's Method).

doi:10.1371/journal.pone.0116309.g005

expression levels are not globally identical to the ones of PA, or to any of the other diploids. So, the morphological similarity between PAA allotriploids and PA allodiploids is not due to strictly conserved mRNA levels between these genomotypes. On the other hand, the similarities may be at the relative transcriptional contribution of each genome type (P and A alleles) to the overall expression level of each gene, regardless of the total expression level of each gene on each hybrid. Other alternative to the above stated, would be the occurrence of conserved protein levels and identities between 2n and 3n hybrids [39]. So, the regulation between PA and PAA at the translational and post-translational levels and the allele specific contribution to the overall expression of each gene should be investigated in a near future.

However, at the transcriptional level, though we ruled out the hypotheses of a global strict full "functional diploidization" of triploids, the majority of transcripts is less represented in PAA then it is in PA and AA genomotypes. One could expect that the level of gene expression would change proportionally to the ploidy variation [40] and not the opposite. But, as an allopolyploid, *S. alburnoides* combines "ploidy rise" with hybridization and so, the effects associated to hybridization have to be considered. The pre-existing differences between the expression profiles of the parental genomotypes and also unpredictable effects of the complexities of an inter-genomic gene expression regulation are manifesting in the expression profiles of the *S. alburnoides* hybrids. We observed that the comparisons of both hybrids with PP are consistent with the comparative profile of AA vs PP, where the vast majority of transcripts are less represented in PP. Additionally, we observe a higher amount of transcripts that are less expressed in PAA then in PA. The observed "overcompensation" supports that between triploid and diploid hybrids, dosage compensation occurs, but it is not accurate. We can speculate that in PAA allotriploids the unbalanced genomic contribution and/or a faulty interaction between the different genomes might have to be compensated through allele specific expression regulation [12], and it may well be that the expression level of each allele might be quite variable and adaptable. So, as in plants and invertebrates, also in the *S. alburnoides* allopolyploid complex there is disruption, due to hybridization and anorthoploidy (odd ploidy) [41] of the quantitative assumptions of additivity, that are usually valid in the case of most homogenomic diploids and autopolyploids [9].

Another hypothesis we can put forward to explain the absence of a positive correlation between copy number and mRNA amount is that the expression level profiles of PAA and PA genomotypes, may be influenced by differences in cell size, which are expected to exist between individuals of different ploidy levels [42]. This hypothesis was not yet explored within the *S. alburnoides* complex, and is barely investigated in other organisms [43].

Additivity

In any hybrid, gene expression is under the influence of divergent genomes, so new qualitative and quantitative gene expression networks are expectedly established, resulting from the interactions of the divergent alleles. In *S. alburnoides*, we observed that for most genes this expression level divergence from the parental genomotypes was not achieved by averaging the parental allelic contributions.

The analysis for additive expression showed that this occurs only for a subset of genes, both in diploid and triploid hybrids of *S. alburnoides* (Fig. 4). The occurrence of non-additive gene expression in hybrids and allopolyploids has been extensively reported in plants, for example in maize [44], rice [45] and *Arabidopsis* [46]. Also in animals as oysters [47] and *Drosophila* [13] the topic was explored and conclusions extended to the animal kingdom. However, for vertebrates, genome wide quantitative gene expression studies were missing.

Interestingly, the results of the available studies on plants and invertebrates are not all coincident. Several showed that the majority of genes are expressed additively [48, 49], while other studies found higher levels of nonadditive expression [13, 44]. The causes for these apparent discrepancies are not yet clear [50]. Anyway, the considerable body of data gathered so far shows a possible positive correlation between size of the fraction of the nonadditively expressed genes and the magnitude of heterotic response. Also, increasing the number of diverse genome copies in an allopolyploid, usually leads to increasingly greater magnitudes of heterosis [51]. Nevertheless, there is no consensus about the amount or identity of the nonadditively expressed genes [50, 52]. Concerning the *S. alburnoides* complex, the phenomenon of heterosis has been barely addressed, except for a few comparisons of growth and reproductive traits between diploid and triploid hybrids [53] and a comparative morphometric study [36], where the AA and PP parental genomotypes were not included. From these studies, mostly non-significant differences between diploid and triploid hybrids have been found, except for a marginal longevity increase in triploids. Yet, the PAA genomotype is far more frequent in the natural populations than PA. So, if we consider the number of non-additively expressed transcripts as an indicator of heterosis, PAA *S. alburnoides* are favored since the amount of additively expressed transcripts is higher than in PA genomotype. Also, according the Bateson-Dobzhansky-Muller Model a lower fitness in hybrids might result from a bad interaction between divergent genomes due to the differential capacity of interaction between their proteins [54]. In this light, allopolyploid individuals have better chances to evade this weakness. Allopolyploids have more options to non-additively combine allele-specific regulated expressions and so, have higher chances to achieve an optimized and more functional expression pattern than one achieved merely additively.

Dosage responses across the *S. alburnoides* hybrid transcriptome—Expression level regulation

In our genome wide prospection for gene expression dosage compensation, a genome wide regulatory mechanism that brings all genic activity of triploids to the diploid state, in a "strict diploidization" of triploids was not seen. However, in both juveniles and liver data sets, we found a considerable fraction of all transcripts (29% in juveniles and 15% in livers) that really suffices the most stringent parameter definition for "fully dosage compensated". So, "diploidization" might not mean that all genes are down regulated to the diploid level, but only those that need to be "diploidized" in RNA amount to function correctly. Also, a considerable part of all transcripts (around half) do not have a significantly different representation between diploid and triploid hybrids (SE group), and from the SE transcripts that do not belong to class I, the majority are lower represented in PAA than in PA (class III). In fact, there is only a very small percentage of triploid transcripts that are represented strictly proportionally to gene dosage (class II) or even higher (class IV). Thus, our results show a significant "diploidization" in triploid PAA genomotype, but not as a strictly regulated and fine-tuned phenomenon. That is consistent with a switch-like way to regulate the mRNA concentrations, where transcription is turned "on" or "off", regardless of exact concentrations [55], but within boundaries of similar expression.

The mechanisms involved in this regulation of gene expression in *S. alburnoides* triploids vs diploids are still elusive, but miRNA's were recently pointed [37] as significant regulators for the functional stability of triploidy in the *S. alburnoides* complex [37].

The quantitative PAA/PA gene expression analysis in juveniles and livers are conceptually coincidental so, most conclusions should be valid both at the single tissue gene expression regulation as at the full-body scale. There is also convergence between our study and the ones the

ones that preceded it. For the same *S. alburnoides* genes analyzed in [12] and [38] we inspected the PAA/PA expression obtained in our study. All were place inside the SE group (data not shown), except *amh* and *dmrt1* (that are not expected to fit with the dataset).

Functional context of the *S. alburnoides* genome regulation

In both, juveniles and liver data sets, the SE group is enriched in terms related to the basal biological maintenance of the cells (eg. metabolism), and mostly in ribosome-linked terms. But regardless of the statistical significance for the enrichment (corrected p-value < 0.05), the fold enrichment of each term is approximately 1. So, in the context of global expression, this enrichment may not be meaningful [35] or may indicate that between triploid and diploid hybrids genes with expression within boundaries of similar RNA amount can occur at any quantity without compromising there function. This expression level flexibility may be a strong contributor to overcome the allopolyploid "genomic shock". Also it gives an immediate evolutionary advantage to the (allo)polyploids.

Within the SE group, we looked for functional enrichment in the strictly dosage compensated class (I) but no significant enrichment was found. That reinforces the previous idea, yet the very small class II, composed of genes with expression strictly proportional to gene dosage, presents a significant enrichment in terms associated to multi-subunit complexes, namely ribosomes. However, the detection of dosage sensitivity in genes whose products are part of multi-subunit complexes is in accordance with the gene balance hypothesis [40, 52], which posits that changes in the stoichiometry of the individual subunits would be deleterious.

The functional enrichment of the DE gene group may shed some light on phenotypic differences between PAA and PA genomotypes. We verified that in both, the liver and juvenile data sets the GO term enrichment is mostly associated to cell surface and to processes intimately linked to the cells membrane. Previously, it was described the same enrichment for differential expression between budding yeasts (*S. cerevisiae*) with different ploidy levels [42]. The authors suggested that the differential gene expression observed between ploidy levels was due to cell size and geometry differences between yeasts of different ploidies and not directly to the gene dosage increase [42]. In addition to yeasts, in many other polyploid organisms, including fish [56, 57, 58], the nucleus and cell volumes expand proportionally to accommodate the enlarged genome of polyploid cells [56]. However, there is a reduction in surface area relative to cell volume [43]. Consequently, the interactions between surface and cytoplasmic signaling, transport of metabolites and the cellular component organization are expected to be affected.

Concluding Remarks

To our knowledge, this is the first study that globally quantitatively and comparatively profiled by mRNA-seq the transcriptomes of diploid and triploid forms of an allopolyploid vertebrate organism.

Our results point towards a certain level of flexibility of expression within a range of mRNA amounts per locus between diploid and triploid hybrids of the *S. alburnoides* complex. For these allotriploids, gene expression levels are similar to the ones of allodiploids but are neither genome wide strictly diploidized nor strictly proportional to gene dosage. The occurrence of a non-fine-tuned expression regulation at the transcription level might be a key factor for the evolutionary success of allopolyploids. Similar to nucleotide sequence variation, variability at the mRNA expression levels may be also a source for regulatory adaptation to selective pressures. Moreover, the evolution of new functions and subfunctionalizations from redundant genes, that are well known to occur in the allopolyploid situation, are probably facilitated in a context of expression dosage plasticity.

In conclusion, this work illustrates how a successful allopolyploid vertebrate transcriptionally deals with the genomic stress derived from hybridization and polyploidy and may shed some light on important features of genome evolution in allopolyploids.

Supporting Information

S1 Table. Output Statistics of sequencing for juveniles' data set.
(DOCX)

S2 Table. Sequencing and mapping statistics for livers data set.
(DOCX)

S3 Table. Statistics of assembly quality for juveniles' data set.
(DOCX)

S4 Table. Summary of annotation results for juveniles' data set (cds information).
(DOCX)

S5 Table. qRT-PCR primers.
(XLSX)

S6 Table. Similarly expressed transcripts (SE) between each pair of 3n vs 2n *S. alburnoides* genomotypes, both in juveniles and liver data sets.
(DOCX)

S7 Table. Significantly differently expressed (DE) transcripts between each pair of 3n vs 2n S. alburnoides genomotype, both in juveniles and liver data sets.
(DOCX)

Acknowledgments

We thank to Susanne Kneitz and to Ana Rita Grosso for helping with the bioinformatics data treatments and discussions.

Author Contributions

Conceived and designed the experiments: MMC MS IM. Performed the experiments: IM. Analyzed the data: IM MPM. Contributed reagents/materials/analysis tools: MMC. Wrote the paper: IM.

References

1. Wang XY, Shi XL, Hao B L, Ge S, Luo JC (2005) Duplication and DNA segmental loss in the rice genome: implications for diploidization. New Phytol 165:937–946. PMID: 15720704
2. Osborn TC, Pires JC, Birchler JA, Auger DL, Chen ZJ, et al. (2003) Understanding mechanisms of novel gene expression in polyploids. Trends Genet 19:141–7. PMID: 12615008
3. Xu Y, Zhao Q, Mei S, Wang J (2012) Genomic and transcriptomic alterations following hybridisation and genome doubling in trigenomic allohexaploid *Brassica carinata* × *Brassica rapa*. Plant Biol doi: 10.1111/j.1438-8677.2011.00553.x
4. Mable BK, Alexandrou MA, Taylor MI (2011) Genome duplication in amphibians and fish: an extended synthesis. J Zool 284:151–182.
5. Alves MJ, Coelho MM, Collares-Pereira MJ (2001) Evolution in action through hybridisation and polyploidy in an Iberian freshwater fish: a genetic review. Genetica 111: 375–385. PMID: 11841181
6. Hotz H, Semlitsch RD, Gutmann E, Guex GD, Beerli P (1999) Spontaneous heterosis in larval life-history traits of hemiclonal frog hybrids. PNAS 96:2171–2176. PMID: 10051613

7. Stöck M, Lamatsch DK, Steinlein C, Epplen JT, Grosse WR, et al. (2002) A bisexually reproducing all-triploid vertebrate. Nat Genet 30:325–8. PMID: 11836500

8. Lampert KP, Schartl M (2008) The origin and evolution of a unisexual hybrid: Poecilia formosa. Philos Trans R Soc Lond B Biol Sci 363:2901–2909. doi: 10.1098/rstb.2008.0040 PMID: 18508756

9. Ritz CM, Köhnen I, Groth M, Theissen G, Wissemann V (2011) To be or not to be the odd one out—allele-specific transcription in pentaploid dogroses (Rosa L. sect. Caninae (DC.) Ser). BMC Plant Biol 23:11–37.

10. Feldman M, Levy AA, Fahima T, Korol A (2012) Genomic asymmetry in allopolyploid plants: wheat as a model. J Exp Bot 63:5045–59. doi: 10.1093/jxb/ers192 PMID: 22859676

11. Mable BK (2013) Polyploids and hybrids in changing environments: winners or losers in the struggle for adaptation? Heredity 110:95–96. doi: 10.1038/hdy.2012.105 PMID: 23321773

12. Pala I, Coelho MM, Schartl M (2008) Dosage compensation by gene-copy silencing in a triploid hybrid fish. Curr Biol 18:1344–1348. doi: 10.1016/j.cub.2008.07.096 PMID: 18771921

13. McManus CJ, Coolon JD, Duff MO, Eipper-Mains J, Graveley BR, Wittkopp PJ (2010) Regulatory divergence in *Drosophila* revealed by mRNA-seq. Genome Res 20:816–25. doi: 10.1101/gr.102491.109 PMID: 20354124

14. Fraser BA, Weadick CJ, Janowitz I, Rodd FH, Hughes KA (2011) Sequencing and characterization of the guppy (*Poecilia reticulata*) transcriptome. BMC Genomics 12:202. doi: 10.1186/1471-2164-12-202 PMID: 21507250

15. Schwarz EM, Kato M, Sternberg PW (2012) Functional transcriptomics of a migrating cell in *Caenorhabditis elegans*. PNAS 109:16246–51. doi: 10.1073/pnas.1203045109 PMID: 22991463

16. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63. doi: 10.1038/nrg2484 PMID: 19015660

17. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 18:1509–17. doi: 10.1101/gr.079558.108 PMID: 18550803

18. Kakumanu A, Ambavaram MM, Klumas C, Krishnan A, Batlang U, et al. (2012) Effects of drought on gene expression in maize reproductive and leaf meristem tissue revealed by RNA-Seq. Plant Physiol 160:846–67. doi: 10.1104/pp.112.200444 PMID: 22837360

19. Collares-Pereira MJ, Matos I, Morgado-Santos M, Coelho MM (2013) Natural pathways towards polyploidy in animals: the *Squalius alburnoides* fish complex as a model system to study genome size and genome reorganization in polyploids. Cytogenet Genome Res 140:97–116. doi: 10.1159/000351729 PMID: 23796598

20. Pittman K, Yúfera M, Pavlidis M, Geffen AJ, Koven W, et al. (2013) Fantastically plastic: fish larvae equipped for a new world. Rev Aquacult 5:S224–S267.

21. Grimplet J, Deluc LG, Tillett RL, Wheatley MD, Schlauch KA, et al. (2007) Tissue-specific mRNA expression profiling in grape berry tissues. BMC Genomics. 8:187. PMID: 17584945

22. Lovén J, Orlando DA, Sigova AA, Lin CY, Rahl PB, Burge CB, Levens DL, Lee TI, Young RA (2012) Revisiting global gene expression analysis. Cell 151(3):476–82. doi: 10.1016/j.cell.2012.10.012 PMID: 23101621

23. Coate JE, Doyle JJ (2010) Quantifying whole transcriptome size, a prerequisite for understanding transcriptome evolution across species: an example from a plant allopolyploid. Genome Biol Evol 2:534–46. doi: 10.1093/gbe/evq038 PMID: 20671102

24. Morgado-Santos M, Matos I, Vicente L, Collares-Pereira MJ (2010) Scaleprinting: individual identification based on scale patterns. J Fish Biol 76:1228–32. doi: 10.1111/j.1095-8649.2010.02591.x PMID: 20409174

25. Sousa-Santos C, Robalo JI, Collares-Pereira MJ, Almada VC (2005) Heterozygous indels as useful tools in the reconstruction of DNA sequences and in the assessment of ploidy level and genomic constitution of hybrid organisms. DNA Seq 16:462–7. PMID: 16287626

26. Crespo López ME, Duarte T, Dowling T, Coelho MM (2006) Modes of reproduction of the hybridogenetic fish Squalius alburnoides in the Tejo and Guadiana rivers: An approach with microsatellites. Zoology 109:277-286. PMID: 16989992

27. Christodoulou DC, Gorham JM, Herman DS, Seidman JG (2011) Construction of normalized RNA-seq libraries for next-generation sequencing using the crab duplex-specific nuclease. Curr Protoc Mol Biol Chapter 4, Unit 4.12. doi: 10.1002/0471142727.mb0412s94

28. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644–52. doi: 10.1038/nbt.1883 PMID: 21572440

29. Pertea G, Huang X, Liang F, Antonescu V, Sultana R, et al. (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics 19:651–2. PMID: 12651724

30. Iseli C, Jongeneel CV, Bucher P (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proc Int Conf Intell Syst Mol Biol 138–48.

31. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5:621–8. doi: 10.1038/nmeth.1226 PMID: 18516045

32. Pontes O, Lawrence RJ, Neves N, Silva M, Lee JH, Chen ZJ, Viegas W, Pikaard CS (2003) Natural variation in nucleolar dominance reveals the relationship between nucleolus organizer chromatin topology and rRNA gene transcription in Arabidopsis. PNAS 100(20):11418–23. PMID: 14504406

33. Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method.Methods. 25:402–8. PMID: 11846609

34. ASAB (2012) Guidelines for the treatment of animals in behavioral research and teaching. Animal Behavior 83:301–309.

35. Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4:44–57. doi: 10.1038/nprot.2008.211 PMID: 19131956

36. Cunha C, Bastir M, Coelho MM, Doadrio I (2009) Body shape evolution among ploidy levels of the *Squalius alburnoides* hybrid complex (Teleostei, Cyprinidae). J Evolution Biol 22:718–28. doi: 10.1111/j.1420-9101.2009.01695.x PMID: 19320794

37. Inácio A, Pinho J, Pereira PM, Comai L, Coelho MM (2012) Global analysis of the small RNA transcriptome in different ploidies and genomic combinations of a vertebrate complex—the Squalius alburnoides. PLoS One 7:e41158. doi: 10.1371/journal.pone.0041158 PMID: 22815952

38. Pala I, Schartl M, Brito M, Malta Vacas J, Coelho MM (2010) Gene expression regulation and lineage evolution: the North and South tale of the hybrid polyploid Squalius alburnoides complex. Proc Biol Sci 277:3519–25. doi: 10.1098/rspb.2010.1071 PMID: 20554543

39. Vogel C (2013) Evolution—Protein expression under pressure. Science 342:1052–3. doi: 10.1126/science.1247833 PMID: 24288321

40. Birchler JA, Veitia RA (2007) The gene balance hypothesis: From classical genetics to modern genomics. Plant Cell 19:395–402. PMID: 17293565

41. Otto SP, Whitton J (2000) Polyploid incidence and evolution. Ann Rev Genet 34:401–437. PMID: 11092833

42. Wu CY, Rolfe PA, Gifford DK, Fink GR (2010) Control of Transcription by Cell Size. Plos Biol 8: e1000523. doi: 10.1371/journal.pbio.1000523 PMID: 21072241

43. Marguerat S, Bähler J (2012) Coordinating genome expression with cell size. Trends Genet 28:560–5. doi: 10.1016/j.tig.2012.07.003 PMID: 22863032

44. Auger DL, Gray AD, Ream TS, Kato A, Coe EH Jr, Birchler JA (2005) Nonadditive Gene Expression in Diploid and Triploid Hybrids of Maize. Genetics 169: 389–397. PMID: 15489529

45. He G, Zhu X, Elling AA, Chen L, Wang X, et al. (2010) Global epigenetic and transcriptional trends among two rice subspecies and their reciprocal hybrids. Plant Cell 22:17–33. doi: 10.1105/tpc.109.072041 PMID: 20086188

46. Miller M, Zhang C, Chen ZJ (2012) Ploidy and Hybridity Effects on Growth Vigor and Gene Expression in *Arabidopsis thaliana* Hybrids and Their Parents. G3 (Bethesda) 2:505–13. doi: 10.1534/g3.112.002162 PMID: 22540042

47. Hedgecock D, Lin JZ, DeCola S, Haudenschild CD, Meyer E, et al. (2007) Transcriptomic analysis of growth heterosis in larval Pacific oysters (*Crassostrea gigas*). PNAS 104:2313–2318. PMID: 17277080

48. Swanson-Wagner RA, Jia Y, DeCook R, Borsuk LA, Nettleton D, et al. (2006) All possible modes of gene action are observed in a global comparison of gene expression in a maize F1 hybrid and its inbred parents. PNAS 103:6805–6810. PMID: 16641103

49. Guo M, Rupe MA, Yang X, Crasta O, Zinselmeier C, et al. (2006) Genome-wide transcript analysis of maize hybrids: allelic additive gene expression and yield heterosis. Theor Appl Genet 113:831–845. PMID: 16868764

50. Stupar RM, Gardiner JM, Oldre AG, Haun WJ, Chandler VL, et al. (2008) Gene expression analyses in maize inbreds and hybrids with varying levels of heterosis. BMC Plant Biol 8:33. doi: 10.1186/1471-2229-8-33 PMID: 18402703

51. Birchler JA, Yao H, Chudalayandi S, Vaiman D, Veitia RA (2010) Heterosis. Plant Cell 22:2105–12. doi: 10.1105/tpc.110.076133 PMID: 20622146

52. Birchler JA, Veitia RA (2012) Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. PNAS 109:14746–53. doi: 10.1073/pnas.1207726109 PMID: 22908297

53. Ribeiro F, Cowx IG, Tiago P, Filipe AF, Moreira Da Costa L, et al. (2003) Growth and reproductive traits of diploid and triploid forms of the *Squalius alburnoides* cyprinid complex in a tributary of the Guadiana River, Portugal. Archiv fur hydrobiologie. 156:471–484.

54. Chang PL, Dilkes BP, McMahon M, Comai L, Nuzhdin SV (2010) Homoeolog-specific retention and use in allotetraploid *Arabidopsis suecica* depends on parent of origin and network partners. Genome Biol 11:R125. doi: 10.1186/gb-2010-11-12-r125 PMID: 21182768

55. Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S (2006) Stochastic mRNA synthesis in mammalian cells. PLoS Biol 4:e309. PMID: 17048983

56. Small SA, Benfey TJ (1987) Cell size in triploid salmon. J Exp Zool 241:339–342.

57. Beyea MM, Benfey TJ, Kieffer JD (2005) Hematology and stress physiology of juvenile diploid and triploid shortnose sturgeon (*Acipenser brevirostrum*). Fish Physiol Biochem 31:303–313.

58. Gao Z, Wang W, Abbas K, Zhou X, Yang Y, et al. (2007) Haematological characterization of loach *Misgurnus anguillicaudatus*: comparison among diploid, triploid and tetraploid specimens. Comp Biochem Physiol A Mol Integr Physiol 147:1001–8. PMID: 17466553

CHAPTER 4

Supplementary data

Table S1. Output Statistics of sequencing for juveniles' data set.

Samples	Total Raw Reads	Total Clean Reads	Total Clean Nucleotides (nt)	Q20 percentage	N %	GC %
juv_AA	86.210.418	80.706.646	7.263.598.140	98.18%	0.00%	49.49%
juv_PA	87.256.908	82.315.202	7.408.368.180	98.18%	0.01%	49.68%
juv_PAA	84.446.254	79.595.526	7.163.597.340	98.16%	0.00%	49.31%

Table S2. Sequencing and mapping statistics for livers data set.

Code	Total reads*	Mapped reads	Properly paird (%)	Fragments	Singletons (%)
liv-AA	55691260	32784631 (59%)	17240502 (31%)	12521017	7742597 (13,9%)
liv-PP	45463238	29025241 (64%)	16141464 (35%)	11234018	6557178 (14,5%)
liv-PA	41955914	23366703 (56%)	113258036 (27%)	861058	6146187 (14,7%)
liv-PAA	56266096	31553986 (56%)	15357742 (27)%)	11574116	8405754 (14,9)
* also QC-passed reads					

Table S3. Statistics of assembly quality for juveniles' data set.

Sample		Total Number	Total Length (nt)	Mean Length (nt)	N50	Total Consensus Sequences	Distinct Clusters	Distinct Singletons
juv_AA	Contig	141,478	64,387,301	455	1008	-	-	-
juv_PAA		176,458	61,061,715	346	641	-	-	-
juv_PA		165,484	64,691,831	391	825	-	-	-
juv_AA	Unigene	89,668	75,011,295	837	1543	89,668	19,292	70,376
juv_PAA		96,276	62,638,402	651	1079	96,276	24,034	72,242
juv_PA		94,919	71,915,349	758	1366	94,919	26,222	68,697
All*		92,137	91,274,231	991	1731	92,137	36,870	55,267
* The contigs from the 3 juveniles libraries								

Table S4. Summary of annotation results for juveniles' data set (cds information).

Sequence File	NR	SwissProt	KEGG	COG	ALL
juv_AA	51,173	44,715	36,636	14,849	71,979
juv_PA	53,067	45,830	37,219	14,664	75,095
juv_PAA	54,769	47,333	37,663	13,871	76,641
All*	52,460	46,149	38,361	16,967	74,463
*joining the 3 juveniles libraries					

Table S5. qRT-PCR primers.

Primer IDs				Primer sequences[2]		Primer binding sites		
Target Genes	Fwd	Rev	Specificity[1]	Fwd	Rev	Fwd	Rev	Product size (bp)
eef1a	eef1a PF2qDNA	ef1a PR2 qDNA	gDNA	GTTTTCAGGTTTTAATTGGCATT	ACCGCTAGCATTACCCTCCT	intron 4-5	exon 5	95
rpl8	rpl8_PF1_qDNA	rpl8_PR1_qDNA	gDNA	GCAAGCAAACATCCCAGTCT	TTCCAGACAGCAGACAATGG	intron3-4	exon 4	137
actb2	actb2_PF1_qDNA	actb2_PR1_qDNA	gDNA	GGATRAATAGWTTTGGGCTGA	CCTTCTGTCCCATACCAACC	intron 2-3	exon 3	144
eef1a	eef1a PF2 qcDNA	eef1a PR 2 qcDNA	cDNA	TCTTGATGCCCTGGATGC	CAGTTCCAATACCTCCAATTTTGT	exon 5	exon 5/6	101
rpl8	rpl8_PF2_qcDNA	rpl8_PR2_qcDNA	cDNA	CAAGAAAGCCCAGCTGAACA	GGATCCAGATGGAAGCTTGA	exon 3/4	exon 4/5	190
actb2	actb2_PF1_qcDNA	actb2_PR1_qcDNA	cDNA	ACATCAGGGTGTCATGGTTG	TCCATATCGTCCCAGTTGGT	exon 2/3	exon 3	99
rpsa	rpsa_PF3_ qcDNA	rpsa icDNA R2	cDNA	GTGACTGATCCTCGTGCTGA	CACAGAGTGGGGACCTTTGT	exon 4	exon 4/5	144
pabpc1a	pabpc1a_PF3_qcDNA	pabpc1a_PR3_qcDNA	cDNA	CAGCCAGTACATGCAGAGGA	ATTCTGAGCCTGTGGAATGG	exon 8	exon 8/9	118
rpl35	rpl35_PF1_qcDNA	rpl35_PR1_qcDNA	cDNA	CCATCGAGAAAAATGGCAAA	GTCATCCAGCTGTTTCAGCA	exon 1/2	exon 2	81

[1]Primers exclusively amplify spliced complementary DNA (cDNA) or exclusively amplify unspliced genomic DNA (gDNA)

[2]underlined portion of sequences corresponds to the second gene region in the "Primer binding sites" column

Table S6. Similarly expressed transcripts (SE) between each pair of 3n vs 2n *S. alburnoides* genomotypes, both in juveniles and liver data sets.

	Comparisons	SE	SE class	SE *per* class	% of SE	% of total
juveniles	***PAA/PA***	58076 **(64%)**	**I**	26376	45%	29%
			II	9935	17%	11%
			III	19672	34%	22%
			IV	2093	4%	2%
	PAA/AA	49778 **(55%)**	**I**	20925	42%	23%
			II	7148	14%	8%
			III	20087	40%	22%
			IV	1618	3%	2%
livers	***PAA/PA***	10068 **(44%)**	**I**	3508	35%	15%
			II	1308	13%	6%
			III	4947	49%	21%
			IV	305	3%	1%
	PAA/AA	9075 **(38%)**	**I**	3823	42%	16%
			II	1359	15%	6%
			III	3553	39%	15%
			IV	340	4%	1%
	PAA/PP	9013 **(41%)**	**I**	3473	39%	16%
			II	2970	33%	13%
			III	990	11%	4%
			IV	1580	17%	7%

Total numbers and percentages of SE unigenes (juveniles) and mapped genes (livers) and total numbers and percentages of SE's *per* expression class.

Table S7. Significantly differently expressed (DE) transcripts between each pair of 3n vs 2n S. alburnoides genomotype, both in juveniles and liver data sets.

	Comparisons	DE		
juveniles	***PAA/PA***	20468 **(22.5%)**	Up	6813 **(7,5%)**
			Down	13655 **(15%)**
	PAA/AA	30024 **(33%)**	Up	12504 **(14%)**
			Down	17520 **(19%)**
liver	***PAA/PA***	195 **(0.83%)**	Up	41 **(0.17%)**
			Down	154 **(0.65%)**
	PAA/AA	261 **(1.1%)**	Up	64 **(0.27%)**
			Down	197 **(0.83%)**
	PAA/PP	52 **(0.23%)**	Up	39 **(0.18%)**
			Down	13 **(0.06%)**

Total numbers and percentages of differently expressed (DE) unigenes in juveniles and mapped genes in livers. DE's were divided in two groups: significantly higher expressed in PAA compared to PA (DEH), and significantly lower expressed in PAA compared to PA (DEL).

CHAPTER 5

Gene copy silencing and DNA methylation in natural and artificially produced allopolyploid fish

Matos I, Coelho MM, Schartl M. Gene copy silencing and DNA methylation in natural and artificially produced allopolyploid fish. J Exp Biol. 219:3072-3081. (2016)

© 2016. Published by The Company of Biologists Ltd | Journal of Experimental Biology (2016) 219, 3072-3081 doi:10.1242/jeb.140418

RESEARCH ARTICLE

Gene copy silencing and DNA methylation in natural and artificially produced allopolyploid fish

Isa M. N. Matos[1,2,*], Maria M. Coelho[1] and Manfred Schartl[2,3,4]

ABSTRACT

Allelic silencing is an important mechanism for coping with gene dosage changes in polyploid organisms that is well known in allopolyploid plants. Only recently, it was shown in the allotriploid fish *Squalius alburnoides* that this process also occurs in vertebrates. However, it is still unknown whether this silencing mechanism is common to other allopolyploid fish, and which mechanisms might be responsible for allelic silencing. We addressed these questions in a comparative study between *Squalius alburnoides* and another allopolyploid complex, the Amazon molly (*Poecilia formosa*). We examined the allelic expression patterns for three target genes in four somatic tissues of natural allo-anorthoploids and laboratory-produced tri-genomic hybrids of *S. alburnoides* and *P. formosa*. Also, for both complexes, we evaluated the correlation between total DNA methylation level and the ploidy status and genomic composition of the individuals. We found that allelic silencing also occurs in other allopolyploid organisms besides the single one that was previously known. We found and discuss disparities within and between the two considered complexes concerning the pattern of allele-specific expression and DNA methylation levels. Disparities might be due to intrinsic characteristics of each genome involved in the hybridization process. Our findings also support the idea that long-term evolutionary processes have an effect on the allele expression patterns and possibly also on DNA methylation levels.

KEY WORDS: Allelic silencing, Allopolyploidy, DNA methylation, Freshwater fish, *Poecilia formosa*, *Squalius alburnoides*

INTRODUCTION

In allopolyploid organisms, ancestral homologous alleles that diversified during evolution, designated 'homoeologs', are brought together again in one individual. Consequently, a successful allopolyploidization process requires the reconciliation of two or more sets of diverged genomes in the same nucleus (Feldman et al., 2012). Importantly, the regulatory interactions between genomes must be stabilized as the increased ploidy level and increased heterozygosity lead to gene redundancy, altered gene dosage and altered relationships within and between loci (Feldman et al., 2012; Yoo et al., 2013). These features make allopolyploid plants and animals exciting objects for understanding the molecular mechanisms of gene regulation in an evolutionary context.

However, studies of the different aspects of allopolyploidy are strongly biased towards plant models (Mable, 2003; Stöck and Lamatsch, 2013). A few years ago, data on the mechanisms underlying gene expression regulation and the dynamics of genome-specific expression in vertebrate allopolyploids were almost absent. Pala et al. (2008) reported for the first time a regulation mechanism of 'functional diploidization' involving gene-copy silencing in an allopolyploid vertebrate, the *S. alburnoides* complex. *Squalius alburnoides* is a hybridogenetic fish that resulted from a cross of a *Squalius pyrenaicus* female (contributing the *p* genome) with an *Anaecypris*-like male (contributing the *a* genome) (Alves et al., 2001) (Fig. S1A). It emerged between 1.4 million years ago (MYA) (Cunha et al., 2004) and less than 0.7 MYA (Sousa-Santos et al., 2007). In present days the complex comprises several ploidy levels and genomic compositions distributed across the Iberian Peninsula (Alves et al., 2001; Collares-Pereira et al., 2013). Taking advantage of the hybrid status of *S. alburnoides*, genome-specific sequence differences were used to determine the contribution of each parental genome to the overall expression of loci individually analyzed in diploid and triploid hybrid individuals (Pala et al., 2008). Results showed that in most triploid *S. alburnoides* of *paa* genome composition, which is the most common form in Iberian southern river basins, for several loci and in different tissues the unpaired minority genome, the *p* haplome, was not contributing to the overall expression, whereas it was contributing to expression in other tissues. Also, the observed allelic expression patterns were different between genes and between different tissues for the same gene. This indicated a most extreme case of homoeolog expression bias (Grover et al., 2012), namely, allele silencing (AS). Therefore, in *S. alburnoides*, the problem of keeping the balance of the expression regulatory networks in an uneven-numbered genomic context might have been solved by AS. These observations were in accordance with gene regulation phenomena already reported in polyploid plants, which showed patterns of differential expression according to organs (Adams et al., 2003) and non-additiveness of expression following gene copy rise (Auger et al., 2005; Wang et al., 2006).

However, it remained unclear whether the silencing mechanism reported for triploid *S. alburnoides*, which is very frequent among both natural and synthesized allopolyploid plants (Adams et al., 2003), was also a common mechanism in allopolyploid vertebrates. A further restriction for generalization is that the allotriploid *S. alburnoides* analyzed so far were all carriers of a duplicated genomic set from one parental species and an unpaired genomic set from another parental species: *paa* and *ppa* in southern populations, and *cca* and *caa* in northern populations, where *S. pyrenaicus* is absent and is replaced by *Squalius carolitertii* (contributing the *c* genome) (Pala et al., 2008, 2010). This situation did not allow the exclusion of monoallelic expression in those cases where the minority genome was not expressed.

[1]Centre for Ecology, Evolution and Environmental Changes, Faculdade de Ciências, Universidade de Lisboa, Lisboa 1749-016, Portugal. [2]Department of Physiological Chemistry, Biocenter, University of Würzburg, Würzburg 97078, Germany. [3]Comprehensive Cancer Center Mainfranken, University Clinic Würzburg, Würzburg 97078, Germany. [4]Texas Institute for Advanced Study and Department of Biology, Texas A&M University, College Station, TX 77843, USA.

*Author for correspondence (immatos@fc.ul.pt)

I.M.N.M., 0000-0002-3725-6742

Received 13 March 2016; Accepted 19 July 2016

Journal of Experimental Biology

RESEARCH ARTICLE Journal of Experimental Biology (2016) 219, 3072-3081 doi:10.1242/jeb.140418

List of abbreviations

5-mC	5-methylcytosine
AS	allelic silencing
HK	housekeeping (gene)
SNP	single nucleotide polymorphism
TF	transcription factor
TFBS	transcription factor binding site
TGH	tri-genomic hybrid

So far, the molecular mechanism responsible for AS in the *S. alburnoides* complex is unknown. A reasonable explanation could be an epigenetic regulation. CpG methylation has long been recognized as a gene expression regulation mechanism by which genes can be silenced by methylation and turned on by demethylation (Martienssen and Colot, 2001). In allopolyploid plants, it is known that among the dramatic genome reconfigurations that can be induced by allopolyploidy, epigenetic changes can play a major role (Wang et al., 2014). However, epigenetic research in (allo)polyploid animals is scarce (Xiao et al., 2013; Covelo-Soto et al., 2015).

To answer these questions and contribute to a better understanding of gene expression regulation in a genomic context of raised ploidy and heterozygosity, we performed a comparative study between *S. alburnoides* and another allopolyploid complex, the Amazon molly (*Poecilia formosa*). *Poecilia formosa* is a unisexual all-female species that originated from a hybridization event between a *Poecilia mexicana limantouri* female (*m* genome) and a *Poecilia latipinna* male (*l* genome) (Lampert and Schartl, 2008) (Fig. S2A), and occurs in the Atlantic drainages, from Rio Tuxpan, Mexico, to South Texas, USA. It reproduces by gynogenesis, thus it depends on sperm from closely related gonochoristic (bisexual) species to trigger embryogenesis of their unreduced diploid eggs (Lampert and Schartl, 2008). Generally, paternal genes do not contribute to the next generation because the paternal pronucleus does not fuse with the unreduced diploid oocyte nucleus. Hence, the vast majority of *P. formosa* are diploid and genetically identical to their mothers. However, in rare cases, the exclusion mechanism fails and paternal introgression occurs (Lampert and Schartl, 2008). In one scenario, small parts of male genetic material are included as microchromosomes (Nanda et al., 2007). In other cases, the sperm nucleus fuses with the oocyte nucleus, resulting in triploid offspring. Such triploids are found in the wild and are true natural allopolyploids having an *mml* genomotype (Fig. S2B). They are fertile and produce all triploid offspring. It has, however, been demonstrated that the formation of such persisting triploid clones is an extremely rare event (Lampert et al., 2005; Schories et al., 2007). These allopolyploidizations were traced back to the evolutionary past of *P. formosa* and have to be considered as ancient events.

The naturally occurring old triploid *P. formosa* (*mml*) are gynogenetically maintained in nature and in the laboratory. On the contrary, triploids that are obtained *de novo* from diploid *P. formosa* as rare introgression cases in laboratory broods (Nanda et al., 1995) do not give rise to stable gynogenetic lines. These *de novo* triploids comprise different genomotypes depending on the parental species used for breeding, including tri-genomic hybrids (TGHs) with *mls* (*P. formosa*, *ml*, with introgressed genome from *P. salvatoris*, *s*) or *mlb* (*P. formosa*, *ml*, with introgressed genome from black molly, *b*) genomic composition (Lamatsch et al., 2010) (Fig. S2C). Such individuals are of great advantage for studying AS in allopolyploids because they offer the opportunity to distinguish all three alleles and evaluate their expression contribution if diagnostic single nucleotide polymorphisms (SNPs) can be found.

To also obtain TGHs of the *S. alburnoides* complex, advantage was taken from the existence of another *Squalius* species, *Squalius aradensis* (*q* genome), which was reported to naturally hybridize with *S. alburnoides* (Sousa-Santos et al., 2006). Thus, triploid hybrids with the *pqa* genomotype can be produced and studied.

In this work, we examined the allelic expression patterns in several somatic organs of diploid and allotriploid *S. alburnoides* and *P. formosa* with particular analyses of TGHs. As a first step towards a mechanistic explanation, we also evaluated the correlation between levels of DNA methylation and the ploidy status and genomic composition of *S. alburnoides* and *P. formosa*.

We show that AS occurs both in *S. alburnoides* and in *P. formosa*. However, we found disparities within and between the two allopolyploid complexes concerning the pattern of allele-specific expression and DNA methylation levels. Our results indicate that long-term evolutionary processes affect allele expression patterns and DNA methylation levels. This study highlights that the relationships between polyploidy, hybridization, methylation and AS are far from linear, and underscores once more the need for further studies in this field.

MATERIALS AND METHODS

Fish samples

Squalius alburnoides (Steindachner 1866) and *S. pyrenaicus* (Günther 1868) were collected from the Almargem stream [29°S; 622,495.24 m E; 4,113,964.49 m N (UTM)], and *S. aradensis* (Coelho, Bogutskaya, Rodrigues & Collares-Pereira 1998) specimens were collected from Arade River basin [29°S; 545,693 m E; 4,133,136 m N (UTM)]. Fish were captured by electrofishing and brought alive to the animal facility of the Faculdade de Ciências da Universidade de Lisboa. Fish were maintained in high-quality glass tanks (30 litres capacity) equipped with filtration units, at 18°C and under a 14 h:10 h light:dark cycle. A *pa S. alburnoides* female and an *S. aradensis* male (previously genotyped) with evident sexual maturation and ready for breeding were used to perform an experimental cross in order to obtain a progeny specifically with *pqa* genotypes. Eggs and sperm were collected from the selected individuals applying gentle pressure to the abdomen and immediately mixed in a Petri dish with water. For 1 year, the progeny was reared constantly at 20°C. Several individuals were genotyped according to Sousa-Santos et al. (2005) in order to confirm the *pqa* genotype of the batch.

Poecilia mexicana limantouri (Jordan & Snyder 1899), *Poecilia latipinna* (Lesueur 1821), *Poecilia salvatoris* (Regan 1907), black molly and *Poecilia formosa* (Girard 1859) individuals were raised and maintained at standard conditions according to Kallman (1975), under a light cycle of 14 h:10 h light:dark. All fish were derived from laboratory stocks of the aquarium of the Biocenter at the University of Würzburg, Germany, that were originally established from fish collected in the wild, except for black molly, which is an ornamental variety of the *P. mexicana*/*P. sphenops* species complex. The strains used in this work are listed in Table S1.

Ethics statement

Fish were captured, handled and euthanized with the approval of the Portuguese National Forest Authority (AFN; fishing credentials nos 53/2013 and 51/2014) and the Biodiversity and Nature Conservation Institute (ICNB; license nos 235/2013/CAPT and 262/2014/CAPT), the Portuguese national authority and relevant

RESEARCH ARTICLE Journal of Experimental Biology (2016) 219, 3072-3081 doi:10.1242/jeb.140418

body concerned with protection of wildlife. The maintenance and use of animals in the animal facility of the Faculty of Science of University of Lisbon (FCUL) had the approval of the Portuguese Directorate-General of Veterinary (DGV), Directorate of Health Services and Animal Protection (DGV-DSSPA) (circular letter no. 99-0420/000/000-9/11/2009).

The selected populations for the fish captures were not imperiled, and sampling was done avoiding depletion of the natural stock. Fish were handled following the recommended ethical guidelines described in the 'Guidelines for the treatment of animals in behavioural research and teaching' (*Animal Behaviour*, 2006, 71, 245–253), and at all times, all efforts were made to minimize fish discomfort. Individuals were submitted to an overdose of the anesthetic MS222 before they were quickly decapitated. Only then were the organs harvested. Fish that were not used were later returned to the collecting site.

All *P. formosa* individuals and fish from parental species used in this study were raised under standard conditions in the aquarium facility of the Biozentrum at the University of Würzburg, where studies were approved by the Institutional Review Board.

Ploidy determination

Fin cells were stained with DAPI as described previously (Lamatsch et al., 2000). At least 10,000 cells were measured per sample. Chicken blood (2.5 pg of DNA per erythrocyte) was used as standard (Vinogradov, 1998).

DNA and RNA extraction

Total genomic DNA was obtained from dissected livers and muscle with a standard phenol/chloroform/isoamyl alcohol (25/24/1) protocol (Blin and Stafford, 1976). DNA was quantified using a Qubit 2.0 Fluorometer (Life Technologies, Carlsbad, CA, USA).

RNA was extracted from dissected livers, eyes, muscle and gills preserved in RNAlater (Ambion, Foster City, CA, USA) at −20°C. Total RNA was extracted using the Tri-Reagent (Ambion) following the supplier's instructions. Contaminant DNA was eliminated by the addition of TURBO DNase (Ambion) followed by purification with phenol/chloroform. Ethanol and glycogen were used to precipitate the RNA. RNA amount and quality evaluation was performed with Nanodrop 1000 (Thermo Fisher Scientific, Waltham, MA, USA) and a 2100 Bioanalyser (Agilent Technologies, Santa Clara, CA, USA).

Sequence analysis and genome-specific expression

From the extracted RNA, first-strand cDNA was synthesized with the RevertAid First Strand cDNA Synthesis Kit (Fermentas, Thermo Fisher Scientific) with oligo dT primers. Primer sequences and amplification conditions for *actb*, *rpl8* and *gapdh* with *Squalius* and *Poecilia* samples are given in Table S2.

In *S. alburnoides*, SNPs between the P and A genomes for the three genes have already been reported (Pala et al., 2008; Matos et al., 2011). For the *S. aradensis* derived Q genome of the *S. alburnoides* complex and for all genomes present in allotriploid *P. formosa*, SNPs were identified in the present study.

PCR products were sequenced and sequences were aligned and compared with Sequencher ver.4 (Gene Codes Corporation, Ann Arbor, MI, USA). Within each of the fish complexes, polymorphic sites between the intervenient genomes were identified.

cDNA samples from adult liver, eye, gill and muscle of *S. alburnoides* and *P. formosa* diploid and triploid natural hybrids and TGHs were used as templates for independent amplifications and direct sequencing of gene products of the three target genes (*actb*, *rpl8* and *gapdh*). Through sequence comparison, on the basis of the identified polymorphic sites between the involved genomes *p*, *a* and *q*, or *m*, *l*, *s* and *b*, the contribution of each genome-specific allele to the overall expression at each of the three target loci was determined.

Global DNA methylation quantification

The percentage of methylated DNA for the genomotypes of each one of the allopolyploid complexes was determined by colorimetric quantification of 5-methylcytosine (5-mC). Three to five specimens were sampled and analyzed independently for each genomotype. One hundred nanograms of DNA of each individual were loaded into each well of the MethylFlash Methylated DNA Quantification Kit (Epigentek, Farmingdale, NY, USA). The protocol and calculations were performed according to the manufacturer's instructions.

In addition, the observed mean methylation level for each genomotype in the hybrids (diploids and triploids) was compared with an expected methylation level, which was calculated by considering that each of the *p*, *a* and *q* genomes in the hybrids would be methylated at the same level as in the non-hybrid situation. The mean methylation level obtained for each parental diploid genomotype (*pp*, *aa* and *qq*) was used to calculate the expected methylation level for each hybrid genomotype [(*pp*/2)+(*aa*/2)+ (*qq*/2)=additive expectation]. Expected additive values for *P. formosa* were calculated accordingly.

The mean observed methylation value (obs) for each hybrid genomotype was divided by its corresponding expected additive value (exp) (Table S3).

Comparative sequence analysis for promoter and CpG island predictions

Sequences for *P. formosa*, *P. mexicana* and *P. latipinna rpl8* (ID: 103134768, 106918910 and 106964237, respectively), *gapdh* (ID: 103136734, 106921370 and 106955760, respectively) and actb (ID: 103153440, 106927995 and 106956540, respectively) were obtained from GenBank. Ensemble84 Amazon molly gene annotations were used to identify exons, introns and untranslated regions. Putative promoter regions were defined as 2000 bp 5′ of the first nucleotide of the first exon (adapted from Farré et al., 2007).

For each gene, sequences were aligned and compared using Bioedit (Hall, 1999) with ClustalW multiple sequence alignment. The putative promoter regions served as templates for the design of degenerated primer pairs that were used to amplify the homoeologous DNA regions from *P. salvatoris* and black molly liver DNA samples. Primer sequences and amplification conditions are given in Table S2.

PCR products were sequenced and all sequences for each gene were aligned as previously.

Several tools were employed to analyze the nucleotide sequence of the putative promoter regions of *rpl8*, *gapdh* and *actb* between *mm*, *ll*, *bb* and *ss* genomes. Identity matrices were obtained with BioEdit. Promoter 2.0 Prediction Server (Knudsen, 1999) and the Gene Promoter Miner (Lee et al., 2012) were used to predict RNA polymerase II (Pol.II) promoters in *Poecilia* DNA sequences. With the Sequence Manipulation Suite – CpG Islands Sequence Analysis option (Stothard, 2000), the occurrence of CpG islands was predicted. Also, with DBCAT (Kuo et al., 2011) the occurrence of CpG islands was investigated as well as the number of CpG per 1 kb within the *mm*, *ll*, *bb* and *ss* sequences.

Journal of Experimental Biology

RESEARCH ARTICLE Journal of Experimental Biology (2016) 219, 3072-3081 doi:10.1242/jeb.140418

RESULTS

Analysis of allele-specific gene expression in triploid *S. alburnoides*

In *S. alburnoides* individuals we analyzed the qualitative pattern of allele-specific contribution for three genes, *actb*, *rpl8* and *gapdh*, in liver, muscle, eye and gill of naturally occurring allotriploids (*paa* genomotype) and laboratory-produced TGHs (*pqa* genomotype).

Several informative SNPs between *p* and *a* alleles for *actb*, *rpl8* and *gapdh* were previously reported (Pala et al., 2008; Matos et al., 2011) and used for this study. When *q* sequences were inspected and compared with *p* and *a* sequences, diagnostic SNPs between them were also identified.

The sequencing of reverse-transcribed PCR products of these three genes from all four organs once again confirmed that in *paa* individuals, AS of *p* is occurring (Table 1). Consistent with previous reports, monoallelic expression of the single *p* allele was not detected.

In contrast, in the TGH hybrids containing the *q* genome, sequencing of the reverse-transcribed PCR products of all three genes revealed no indication of silencing in any of the four analyzed tissues. The observed qualitative pattern of allele usage in the TGH individuals was consistently tri-allelic (Table 1).

Allele-specific expression in triploid *P. formosa*

For naturally occurring *P. formosa* allotriploids (*mml*) and laboratory-produced TGHs (*mlb* and *mls*), the qualitative pattern of allelic-specific contribution in *actb*, *rpl8* and *gapdh* in the liver, muscle, eye and gill was inspected (Table 2). Contrary to what was observed in *S. alburnoides*, in natural triploid *P. formosa* (*mml*) no evidence for AS was obtained.

We then looked at the laboratory-generated TGHs, either with *mlb* or *mls* genomic composition (Table 2). For both types of TGH we clearly detected allele-specific silencing. Moreover, in *mls* TGHs for *gapdh* and *rpl8*, even monoallelic gene expression (silencing of two alleles) was detected.

Global DNA methylation in *S. alburnoides* of different ploidy levels and genomic composition

Allele-specific silencing can be due to an epigenetic mechanism. Therefore, we determined the total amount of 5-mC in total DNA extracts from livers and muscle of natural allodiploid (*pa*), allotriploid (*paa*) and laboratory-produced TGH (*pqa*) *S. alburnoides*, as well as from the parental non-hybrids – *aa*, *pp* and *qq* (Fig. 1A,B). In both liver and muscle samples, there was a significantly higher amount of 5-mC in the *aa* diploids than in all other diploids. We found also that both natural triploids (*paa*) and the TGH triploids (*pqa*) have a similarly high level of 5-mC as the *aa* diploids, and again significantly higher (*t*-test for independent samples, $P>0.05$) than the *pp*, *qq* and *pa* diploids.

Global DNA methylation in *P. formosa* of different ploidy levels and genomic composition

For *P. formosa* we determined the global 5-mC levels in natural allodiploids (*ml*), allotriploids (*mml*), TGHs (*mls* and *mlb*) and in all the parental diploids (*mm*, *ll*, *bb* and *ss*) (Fig. 1C,D). For all *Poecilia* genomotypes, the pattern of 5-mC was consistent between the two analyzed tissues. In both liver and muscle, higher levels of 5-mC were found in the natural diploid and triploid hybrids, while all diploid parental genomotypes (*mm*, *ll*, *bb* and *ss*) and the laboratory-produced TGH (*mlb* and *mls*) displayed a similar low methylation level.

Additivity of global DNA methylation in *S. alburnoides* and *P. formosa* allopolyploid complexes

For each hybrid genomotype we performed a simple relative comparison (ratio) between the mean observed methylation value and an expected methylation level in case of additivity (obs/exp) for a hybrid situation (Table S3). Results show that the genomotypes of both allopolyploid complexes can be separated into two distinct groups. One group is composed of *pa*, *paa*, *pqa*, *mlb* and *mls* genomotypes, with obs<exp, and a second group is composed of *ml* and *mml* genomotypes, with obs>exp (Fig. 2).

Promoter and CpG island prediction of *Poecilia* target genes

We used available genomic sequences of *P. mexicana*, *P. latipinna* and *P. formosa* as templates to isolate and characterize the homoeologous sequences in *P. salvatoris* and black molly. The selected target zones were the 2000 bp 5′ of the first nucleotide of the first exon of *rpl8*, *gapdh* and *actb*. We could amplify between 1100 and 1429 bp for *P. salvatoris* and black molly within these template regions. For each gene, we found, as expected for comparisons between species and/or strains, a high percentage (98–99% for *actb*, 93–99% for *gapdh* and 97–99% in *rpl8*) of positive similarity between *mm*, *ll*, *bb* and *ss* sequences (Table S4).

Within the selected sections for *mm*, *ll*, *bb* and *ss* we could predict for *gapdh* and for *actb* a highly likely promoter region – *gapdh*, from −592 to −294 bp of the first nucleotide of the first exon; and *actb*, from −611 to −296 bp of the first nucleotide of the first exon. Concerning CpG islands, for none of the individual genomes at any of the three genes were any CpG islands predicted with the DBCAT within the defined target zone, but with the Sequence Manipulation Suite a CpG island was found within the defined target zone for *rpl8* (from −498 to −283 bp of the first nucleotide of the first exon) and *actb* (from −1411 to −1204 bp of the first nucleotide of the first exon). Also, we quantified the number of CpG sites per 1 kb within the *mm*, *ll*, *bb* and *ss* sequences (Table S5), but no substantial differences were found between the genomotypes for each gene.

Table 1. Allelic expression pattern of *actb*, *rpl8* and *gapdh* in liver, eye, gill and muscle of *Squalius alburnoides*

			Liver			Eye			Gill			Muscle		
Ploidy level	*n*	Genotype	*actb*	*rpl8*	*gapdh*	*actb*	*rpl8*	*gapdh*	*actb*	*rpl8*	*gapdh*	*actb*	*rpl8*	*gapdh*
2n	2	*pa*	*p+a*	*p+a*	*p+a*	*p+a*	*p+a*	*p+a*	*p+a*	*p+a*	*p+a*	*p+a*	*p+a*	*p+a*
3n	5	*paa*	*p+a*	*p+a*	*p+a*	*p+a*	*p+a*	*p+a*	*p+a*	*p+a*	*p+a*	*p+a*	*p+a*	*p+a*
3n	1	*paa*	***a***	*p+a*	*p+a*	*p+a*	*p+a*	*p+a*	*p+a*	*p+a*	*p+a*	*p+a*	*p+a*	*p+a*
3n	1	*paa*	***a***	*p+a*	*p+a*	*p+a*	*a*	*p+a*	*p+a*	*p+a*	*p+a*	*p+a*	*p+a*	*p+a*
3n	6	*pqa*	*p+q+a*	*p+q+a*	*p+q+a*	*p+q+a*	*p+q+a*	*p+q+a*	*p+q+a*	*p+q+a*	*p+q+a*	*p+q+a*	*p+q+a*	*p+q+a*

Allelic silencing is highlighted in bold.

RESEARCH ARTICLE Journal of Experimental Biology (2016) 219, 3072-3081 doi:10.1242/jeb.140418

Table 2. Allelic expression pattern of *actb*, *rpl8* and *gapdh* in liver, eye, gill and muscle of *Poecilia formosa*

			Liver			Eye			Gill			Muscle		
Ploidy level	*n*	Genotype	*actb*	*rpl8*	*gapdh*	*actb*	*rpl8*	*gapdh*	*actb*	*rpl8*	*gapdh*	*actb*	*rpl8*	*gapdh*
2n	2	*ml*	*m+/*	*m+/*	*m+/*	*m+/*	*m+/*	*m+/*	*m+/*	*m+/*	*m+/*	*m+/*	*m+/*	*m+/*
3n	9	*mml*	*m+/*	*m+/*	*m+/*	*m+/*	*m+/*	*m+/*	*m+/*	*m+/*	*m+/*	*m+/*	*m+/*	*m+/*
3n	1	*mls*	*m+/+(s)*	_	***s***	_	_	***s***	*m+/+(s)*	_	_	*m+/+(s)*	_	_
3n	1	*mls*	***l+s***	*m+/+(s)*	***s***	*m+/+(s)*	*m+/+(s)*	_	***m+s***	_	***s***	***l+s***	*m+/+(s)*	***s***
3n	1	*mls*	*m/s+/*	*m/s+/*	***m+l***	*m/s+/*	*m/s+/*	*m/s+/*	*m/s+/*	*m/s+/*	*m/s*	*m/s+/*	*m/s+/*	*m/s+/*
3n	1	*mls*	*m+/+(s)*	*m/s+/*	*l+s+(m)*	*m+/+(s)*	*m/s+/*	***l+s***	*m+/+(s)*	*m/s+/*	*m+/+s*	*m+/+(s)*	*m/s+/*	*m+/+s*
3n	1	*mls*	*m+/+(s)*	***m/s***	***m+s***	*m+/+(s)*	***m/s***	***m+s***	*m+/+(s)*	***m/s***	***m+s***	*m+/+(s)*	*m/s*	*m+/+s*
3n	1	*mls*	***m+s***	***m/s***	***m+s***	***m+s***	***m/s***	***m+s***	***m+s***	***m/s***	***m+s***	*m+s*	***m***	***m+s***
3n	1	*mls*	*m+/+(s)*	***m/s***	***m+s***	*m+/+(s)*	_	***m+s***	*m+/+(s)*	***m/s***	***m+s***	*m+/+(s)*	***m***	***m+s***
3n	1	*mls*	*m+/+(s)*	*m/s+/*	*m+/+s*	*m+/+(s)*	*m/s+/*	*m/s+/*	*m+/+(s)*	*m/s+/*	*m+/+s*	*m+//s*	*m/s+/*	*m+/+s*
3n	1	*mlb*	_	_	_	_	_	_	*m/b+/*	***m/b***	_	_	_	_
3n	1	*mlb*	*m/b+/*	***m/b***	***m/b***	*m/b+/*	*m/b+/*	***m/b***	*m/b+/*	*m/b+/*	*m/b+/*	*m/b+/*	***m/b***	*m/b+/*
3n	1	*mlb*	*m/b+/*	*m/b+/*	*m/b+/*	*m/b+/*	*m/b+/*	*m/b+/*	*m/b+/*	*m/b+/*	*m/b+/*	*m/b+/*	***m/b***	***m/b***
3n	1	*mlb*	*m/b+/*	_	_	*m/b+/*	*m/b+/*	_	*m/b+/*	*m/b+/*	_	*m/b+/*	_	_
3n	1	*mlb*	*m/b+/*	*m/b+/*	*m+/+(b)*	*m/b+/*	*m/b+/*	*m/b+/*	*m/b+/*	*m/b+/*	*m+/+(b)*	*m/b+/*	*m/b+/*	*m+/+(b)*
3n	1	*mlb*	*m/b+/*	*m/b+/*	*m+/+(b)*	*m/b+/*	*m/b+/*	*m/b+/*	*m/b+/*	*m/b+/*	*m+/+(b)*	*m/b+/*	*m/b+/*	*m+/+(b)*

Allelic silencing is highlighted in bold. /, to be read as 'either or'; (), presence or absence of the allele not unequivocally determined; _, allele expression pattern not assessed.

DISCUSSION

In this work we intended to answer three fundamental questions concerning the mechanism underlying gene expression regulation and the dynamics of genome-specific expression in vertebrate allopolyploids. First, we wanted to explore whether the silencing mechanism reported for natural triploid *S. alburnoides* was common to another allopolyploid vertebrate. Second, we wanted to investigate whether, in an allotriploid condition with increased heterozygosity, one of the three alleles is consistently silent, converting triploids into functional diploids. Third, it was our goal to begin to identify possible mechanisms responsible for allele silencing. Specifically, we wanted to evaluate CpG methylation as a candidate mechanism, but other possibilities have been considered.

Allele-specific silencing in *P. formosa*

In TGH *P. formosa* triploids of *mlb* and *mls* genomic composition, AS was obvious and quite frequent. This shows for the first time that AS is indeed not a unique phenomenon in the *S. alburnoides* complex, but is more widespread. This is in line with earlier findings that the variation in pigmentation phenotypes between TGH of

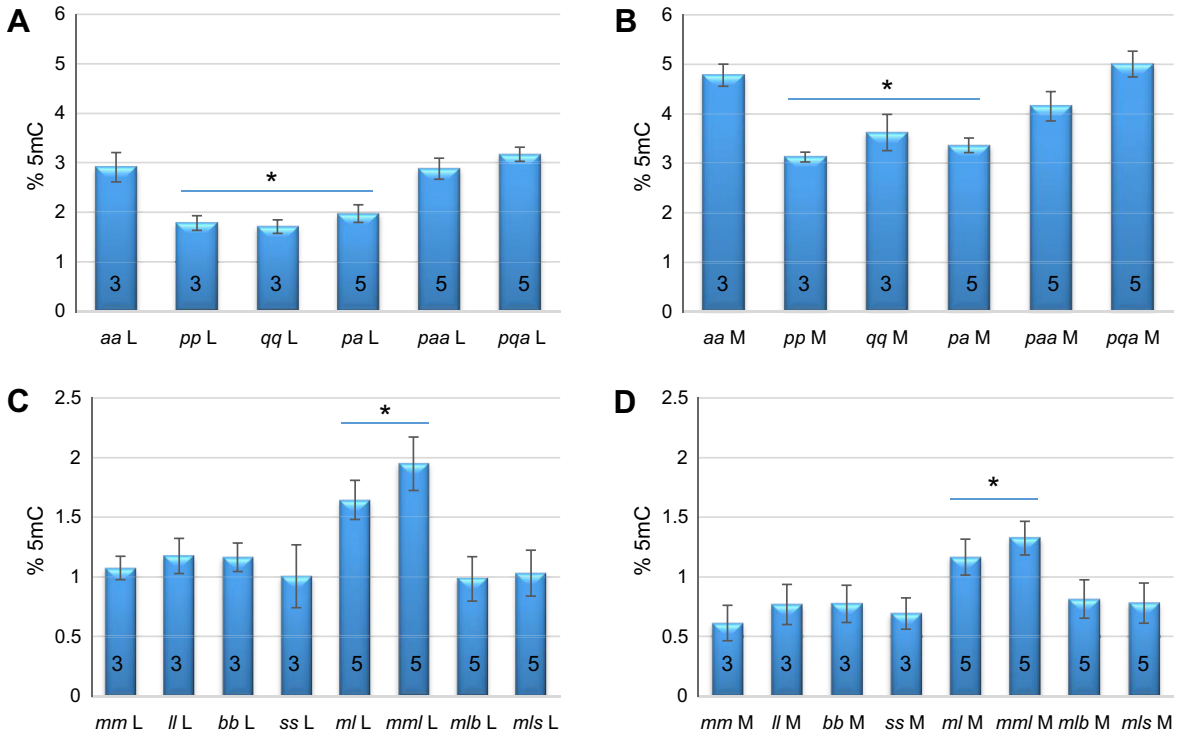


Fig. 1. Levels of global DNA methylation within the *Squalius alburnoides* and *Poecilia formosa* allopolyploid complexes. Global DNA methylation in (A) liver and (B) muscle tissue of several *S. alburnoides* complex intervenient genomotypes. Global DNA methylation in (C) liver and (D) muscle tissue of the *P. formosa* complex intervenient genomotypes. 5-mC, 5-methylcytosine; L, liver; M, muscle. Data are means±s.e.m. *Significant difference (P<0.05) between the underneath genomotypes and all other groups.

RESEARCH ARTICLE Journal of Experimental Biology (2016) 219, 3072-3081 doi:10.1242/jeb.140418

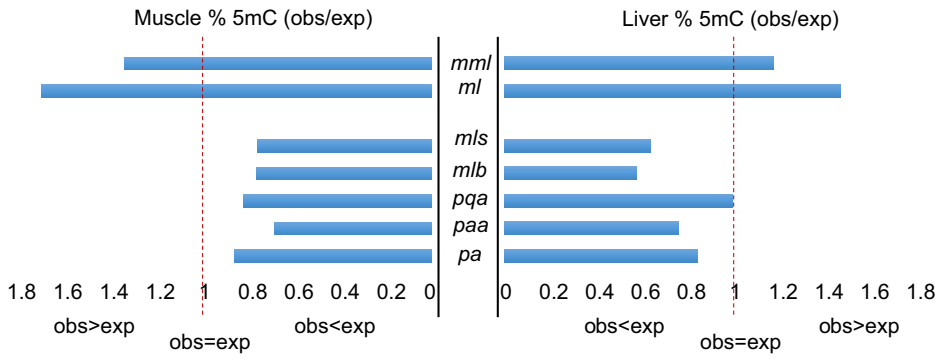


Fig. 2. Additivity of global DNA methylation in *S. alburnoides* and *P. formosa* allopolyploid complexes. Ratio between the mean observed methylation value and an expected methylation level in the case of additivity (obs/exp), in muscle and liver tissues. *pa*, *paa*, *pqa*, *mlb* and *mls* genomotypes present observed 5-mC levels<expected 5-mC levels, whereas *ml* and *mml* genomotypes have observed 5-mC levels>expected 5-mC levels. Dashed red line indicates the position at which observed 5-mC level=expected 5-mC level.

P. formosa individuals may be the consequence of differential contribution of genomes to overall expression (Lamatsch et al., 2010, 2011).

The failure then to detect AS also in the naturally occurring triploid of the *mml* genomic constitution was somehow unexpected as the naturally occurring triploid *P. formosa* were proposed earlier as good candidates where a comparable gene-copy silencing phenomenon like in *S. alburnoides* could occur (Pala et al., 2008). Comparison of expression levels at several allozyme loci between diploid and triploid *P. formosa* revealed them to be indistinguishable quantitatively (Turner et al., 1983), which could be a consequence of AS.

Our failure to detect AS in the naturally occurring *P. formosa* could be due to the following reasons. (1) AS is not random and it is always one of the 'm' alleles that is silenced. This phenomenon would escape our observation because our sequencing chromatograms did not allow for quantitation of peak heights at SNP positions. (2) AS does not occur on a full genomic scale and the three selected genes are not subjected to this phenomenon. However, if there were genome-wide occurrence of AS in triploid *P. formosa*, our study would most likely have been sufficient to detect it. Considering a parsimonious null hypothesis of random inactivation of one of the genomes (neither haplome nor tissue dependent), for each gene and per tissue, 2.7 instances of AS occurrences would be expected (n=9). We analyzed the allele expression pattern in four tissues, so in total per gene, approximately 11 (2.7+2.7+2.7+2.7) 'l' allele silencing occurrences should be seen in our evaluation if this phenomenon exists. If AS is not random and affects only a subset of genes or cell types, more genes and other organs need to be investigated in the future, preferably using transcriptome-wide approaches as recently described by Garcia et al. (2014). (3) AS does not occur at all in the *mml* genomotypes. Although this is a valid assumption in this context, as we did not find AS in naturally occurring allotriploid *P. formosa*, we cannot promptly discard that it does not occur at all. In fact, the occurrence of variegated skin phenotypes presented by some individuals is a strong contra-indicator of this third hypothesis.

The difference between the natural occurring *mml* and the TGH *P. formosa* triploids may be explained by different magnitudes of 'genomic shock'. 'Genomic shock' refers to a series of genomic perturbations at both genetic and epigenetic levels, and has been described in many plant allopolyploid systems (Wang et al., 2014). Some of its most frequent consequences are deviations from expected expression levels and allele specific expression patterns. Also, in plants it has been found that hybridization usually has a greater impact on gene silencing than does genome doubling (Chelaifa et al., 2010; Buggs et al., 2014). Despite both *P. formosa* types having the same ploidy level, the increased diversity of genomes in the TGHs may lead to a higher level of 'genomic shock'. Compared with natural allotriploids, where only two distinct genomes have to be managed, the interactions and simultaneous regulation of three different genomic sets may pose additional challenges with different outcomes. In addition, it has to be considered that some intergenomic combinations are not well tolerated and can lead to hybrid incompatibilities and dysgenesis (Bomblies and Weigel, 2007; Ishikawa and Kinoshita, 2009; Walia et al., 2009; Malone and Hannon, 2009). So, immediate allele-specific expression adjustments in the TGH *P. formosa* may be a necessity to allow for the viability of these organisms.

Absence of AS in TGHs of *Squalius*

Contrary to what was observed in the naturally occurring allotriploid *S. alburnoides*, AS was not observed in any of the analyzed tissues in TGH individuals. It has been previously shown (Pala et al., 2010) that the patterns of gene expression in triploid *S. alburnoides* depend on the genomic contexts brought about by different parental contributions. For instance, the presence of *c* or *p* genomes in allopolyploid *S. alburnoides* biotypes results in substantial difference in genome-specific allele usage in either *paa* or *caa* genomic contexts (Pala et al., 2010). Because the effect of the *q* genome to the overall gene expression in natural occurring *S. alburnoides* of *qaa* and *qqa* genomotypes has never been assessed, the absence of AS in the TGH fish with one *q* haplome is difficult to assess, and the effects of the presence of the *q* genome are difficult to infer. However, we can at least say that the absence of AS in TGH *S. alburnoides* supports the previous conclusion that different genome combinations lead to different mechanisms of how to cope with genomic shock. In contrast, the absence of AS in TGH *Squalius* is not readily explained by the simple reasoning presented for AS occurrence in the TGH *P. formosa*, where we relate the higher genomic shock with the need for AS. This demonstrates the complexity of the phenomenon where two different deviations from normal come together, namely ploidy change and hybridization.

Despite our inability to show AS in the TGH *S. alburnoides*, its occurrence cannot be totally discarded, based on the same considerations presented for the naturally occurring *P. formosa*. So, to fully enlighten this matter, applying a transcriptome-wide approach to *S. alburnoides* would also be desirable.

However, despite new and promising tools that are constantly emerging (Shen et al., 2012a,b, 2013), assessing allele-specific gene expression on a large scale is still a technically challenging problem (Garcia et al., 2014), even more so in species with scarce genomic resources, and as in this case, higher levels of ploidy than diploidy.

RESEARCH ARTICLE Journal of Experimental Biology (2016) 219, 3072-3081 doi:10.1242/jeb.140418

Differences in global DNA methylation between genomotypes

DNA methylation modifications associated with ploidy changes have been studied extensively in plants (Diez et al., 2014). It has been shown that normal function and structure of newly formed polyploid genomes are intimately related with this epigenetic process (Matzke et al., 1999; Salmon et al., 2005; Chen and Ni, 2006; Wang et al., 2014). Also, it is known that methylation impacts directly on gene transcription (Wang et al., 2014; Sehrish et al., 2014). In general, it is assumed that methylated DNA sequences are transcriptionally inactive (Wang et al., 2014). So, one goal of this study was to relate AS occurrence in these fish to the degree of total DNA methylation.

We determined the total amount of DNA methylation in two tissue types (liver and muscle) for all the available genomotypes involved in both allopolyploid complexes. If the AS phenomenon was 5-mC mediated, our hypothesis was that the total methylation level would be higher in those triploid individuals where AS occurs. However, the pattern of global methylation in both the *S. alburnoides* and *P. formosa* allopolyploid complexes does not fit this initial expectation, nor does it help to clarify the different AS patterns between *S. alburnoides* and *P. formosa*. For instance, AS occurs in *P. formosa* TGH, where we identified low levels of methylation compared with naturally occurring diploids and triploids in which AS was not detected. Also, TGH *S. alburnoides*, where no AS was detected, presented similar high levels of methylation as the naturally occurring triploid *S. alburnoides* (*paa* genomotype), where AS has been encountered. So, global methylation levels do not seem to reflect the AS status. This is in line with findings in *Arabidopsis*, where for most of a pool of 77 analyzed genes, expression did not directly correlate with the methylation level (Shen et al., 2012a). In contrast, in *Tragopogon* it was shown that by DNA methylation one homeolog can be completely silenced (Sehrish et al., 2014).

We further observed that the levels of DNA methylation were non-linearly related to the ploidy level in each tested allopolyploid series. Higher ploidy level did not consistently correspond to higher or lower levels of DNA methylation in either of these allopolyploid complexes. Additionally, our results do not show a linear correspondence between higher levels of heterozygosity and higher or lower levels of DNA methylation.

Similar results have been found in an analysis of genomic DNA methylation in several annual herbaceous and woody perennial plants of several ploidy levels (Li et al., 2011). In addition, in a study that investigated DNA methylation changes associated with ploidy in *Salmo trutta*, no evidence of genome-wide methylation differences between diploid and triploid specimens was found (Covelo-Soto et al., 2015). However, in *Cyprinus carpio*×*Carassius auratus* hybrids it was found that hypermethylation was more prominent in the allotetraploids than in the diploid parental individuals (Xiao et al., 2013).

We have determined global methylation levels, but with this broad approach, underlying mechanisms of methylation as effectors at the single-locus scale are diluted. In this sense, investigating differences in 5-mC of promoters of genes presenting AS would be interesting. Methylation of promoters is canonically associated with stable, long-term transcriptional silencing, and one of the reasons is that a transcription factor (TF) is physically prevented from binding to its specific transcription factor binding site (TFBS) if the TFBS is methylated (Zhu et al., 2003; Defossez and Stancheva, 2011). A differential methylation status of CpG sites in the promoter and/or at its surroundings between the different alleles of a gene may lead to differential allelic expression (Kerkel et al., 2008; Sehrish et al., 2014). However, the three target genes focused on in the present study (*rpl8*, *gapdh* and *actb*) are housekeeping (HK) genes. HK genes are expressed in virtually all tissues and across developmental stages and are, in general, exempted from complex transcriptional programs as, for example, the transcriptional programs governing genes involved in responses to external stimuli or in cell differentiation (Farré et al., 2007). In principle, HK genes are activated by default; therefore, the CpG sites around or on the proximal promoter should be unmethylated. Also, contrary to what has been widely reported in other vertebrate organisms, it was found that in zebrafish, methylation and expression were most strongly correlated with regions 10,000 bp upstream and downstream from genes (McGaughey et al., 2014) and not at the proximal promoter sites. So, in the present case, for the specific gene targets on hand, a locus-specific approach did not offer much promise and it was not pursued.

Mechanisms other than DNA methylation may intervene or be responsible for allele expression bias

In any case, mechanisms other than DNA methylation may intervene or be responsible for allele expression bias and AS. For example, an miRNA-linked mechanism has been already identified as a good candidate in the *S. alburnoides* complex (Inácio et al., 2012) and should be similarly investigated for the *P. formosa* complex.

From another angle, in the analysis of the putative promoter regions of *rpl8*, *gapdh* and *actb* of *Poecilia* parental genomotypes, we found a high percentage of positive identity between the sequences. This is an expected result for comparisons within species and/or strains. However, as there is no perfect homology (less than 100% identity), it is conceivable that in the cells of the TGH individuals three different sequences are working simultaneously as promoter of each gene. Conversely, each of these different sequences can work more or less effectively as the docking site for polymerases and transcription factors originated from homoeolog genes. So, another mechanism that may intervene or be responsible for allele expression bias and AS is the strength of the promoter. A promoter can be classified from strong to weak according to its affinity for RNA polymerase and TFs (Li and Zhang, 2014). Thus, the strength of the promoter depends from how closely the promoter sequence resembles the ideal consensus sequences for the docking of polymerase and TFs (Li and Zhang, 2014). For example, in *Escherichia coli* it was observed that several non-consensus bases could have a positive effect on the promoter strength while certain consensus bases have a minimal effect (Kiryu et al., 2005). Also, it was demonstrated in yeasts that variations in the binding sites of TFs between three different strains were responsible for up to 50% of the observed differences in expression (Tirosh et al., 2008). Additionally, a more recent study showed that nucleotides in different regions of promoter sequence have different effects on promoter strength (Li and Zhang, 2014). So, we hypothesize that the conspicuous AS that we encountered in the *P. formosa* TGH may be due to different promoter strengths resulting from the different nucleotidic sequences detected. To support this assumption, a similar analysis for the *S. alburnoides* complex should be performed, and results should show higher levels of identity between the promoter sequences of the parental genomotypes. However, while large-scale annotated genomic data are available for the *P. formosa* complex, no reference genome has yet been produced for the *S. alburnoides* complex, so we could not perform the same analysis.

Journal of Experimental Biology

RESEARCH ARTICLE Journal of Experimental Biology (2016) 219, 3072-3081 doi:10.1242/jeb.140418

'Old' versus '*de novo*' allopolyploids and the effects of long-term evolutionary processes

The analyzed laboratory-bred triploid *P. formosa* individuals with the *mml* genomotype were derived by gynogenesis from natural triploids. In these individuals, the original hybridization (*m*×*l*) and polyploidization (*ml*+*m*) events occurred a long time ago, and are merely clonally propagated at each generation (Lampert and Schartl, 2008). Therefore, we consider them as naturally occurring 'old triploids'. We also analyzed TGH *P. formosa* triploids of *mlb* and *mls* genomic composition that were experimentally produced through specific crosses between *Poecilia* strains and species (Lampert et al., 2007; Lamatsch et al., 2010). We can consider these individuals as '*de novo*' allotriploids, as increases in both ploidy and hybridity happen at the moment of production of each of these TGH individuals.

Inversely to what was observed in the 'old' *P. formosa* triploids, in the '*de novo*' triploids AS was quite frequent and evident. We hypothesize that AS may be an immediate mechanism to cope with the genomic shock. In fact, whenever AS has been detected in vertebrates, it was in individuals that could be considered '*de novo*' triploids. In *S. alburnoides* the reproductive complex is maintained through an intricate network of genetic exchanges and continuous *de novo* hybridizations. Hence, allopolyploidy is established '*de novo*' at the moment of each individual conception. Another example is the laboratory-produced TGH allotriploid medakas (*Oryzias latipes*), where it was found that allele suppression, despite not being abundant, consistently occurred (Garcia et al., 2014).

These examples support the hypothesis that AS may be an immediate mechanism to cope with genomic shock. Consecutively, refined mechanisms operate leading to a stable regulation of the three haplomes. However, we have not found AS in TGH *S. alburnoides*, which are also '*de novo*' allotriploids. This may indicate that AS is not a ubiquitous mechanism to cope with an abrupt increase of ploidy and heterozygosity in fish.

Several studies on allopolyploid plants have also revealed differences between 'old' and 'young' polyploids. The degree of non-additive expression was lower in recent allopolyploids compared with 'older' allopolyploid cotton and coffee genotypes (Flagel and Wendel, 2010). These results suggested that non-additive expression, that is due or related to AS, may increase over time, via selection and modulation of regulatory networks. In another study, results showed that in F1 hybrids and early allopolyploid *Tragopogon miscellus* plants there was activation of allele/homeolog expression in all tissues, eliminating the tissue-specific expression patterns observed in the parental diploids (Buggs et al., 2011). Tissue-specific expression patterns were then reestablished as generations succeeded (Buggs et al., 2011).

In this context, the differences in DNA methylation levels that we observed can also be interpreted. Comparing allotriploids of different evolutionary age, we observed a tendency towards higher DNA methylation levels than expected from additivity in the 'old' hybrids, whereas the opposite tendency was observed in the genomotypes of '*de novo*' hybrids.

In the *S. alburnoides* complex, we found other evidence that long-term evolutionary processes may influence DNA methylation levels. We observed that the percentage of methylated DNA is much higher in the *aa* genomotype than for the other two parental genomotypes (*pp* and *qq*). This may indicate that in individuals of the *aa* genomotype, more genes or alleles are downregulated or inactivated. These increased DNA methylation levels may be related to the fact that both *pp* and *qq* genomotypes exist as independent species (*S. pyrenaicus* and *S aradensis*, respectively), having their own separate evolutionary paths, while an independent species with the *aa* genomotype does not exist. Individuals with the *aa* genomotype, called 'diploid nuclear non-hybrid males of the *S. alburnoides complex*' (Alves et al., 2001), perpetuate only inside the complex by mating with triploid hybrid females (*paa* or *qaa*) (Fig. S1). In each *aa* individual that arises, the nuclear hybrid status is lost and epigenetic changes are likely to occur.

In summary, our results imply that DNA methylation may play some role in the evolution of these vertebrate allopolyploids, probably somehow providing genome stability and reducing the degree of incompatibility that arises from multiple incongruous genomes within the same nucleus. Nevertheless, as in plants, the mechanisms by which all this happens at the whole genomic level (and also at specific sites) seem to be diverse and are still obscure.

Conclusions

With this work, we showed that in vertebrates, AS also occurs in allopolyploid situations besides the previously studied naturally occurring triploid *S. alburnoides*. In *P. formosa* AS was observed quite frequently in two distinct TGH genomic configurations.

We assume that AS is the result of genomic stress, induced by the presence of distinct genomes in the same nucleus. Of note, we found several disparities within and between the two complexes concerning the pattern of allele-specific expression and DNA methylation levels. These differences might be due to the intrinsic characteristics of each genome involved in the hybridization process. Expression silencing or downregulation can result from the interaction between divergent regulatory hierarchies (Riddle and Birchler, 2003) and differential capacity of interaction between proteins or complexes (Comai, 2000; Adams and Wendel, 2004). However, our results also point out that AS is not a ubiquitous mechanism to handle an abrupt increase in ploidy and heterozygosity in fish.

In addition, our findings support the notion that long-term evolutionary processes have an effect on the allele expression patterns and possibly also on DNA methylation levels. Our study highlights the complexity of allopolyploidy at the gene expression regulation level, and that attempts to find a common global mechanism or explanation that fits all allotriploid conditions might fail, as it might not exist.

Acknowledgements

The authors thank Miguel Machado and Miguel Morgado-Santos for their help with fieldwork and critical discussions, and Petra Fischer for ploidy determination of the *Poecilia* specimens.

Competing interests

The authors declare no competing or financial interests.

Author contributions

I.M.N.M. performed the *S. alburnoides* fish capture, carried out the crosses to obtain TGH, performed the experiments, analyzed the data, participated in the design of the study and drafted the manuscript. M.M.C. participated in the design of the study and helped draft the manuscript. M.S. participated in the design of the study, supervised its different components, produced the *P. formosa* TGHs and revised the manuscript. All authors gave final approval for publication.

Funding

This work was supported by Project PTDC/BIA-BIC/110277/2009 to M.M.C. and by a PhD grant (SFRH/BD/61217/2009) to I.M.N.M., both from the Portuguese National Science Foundation, Fundação para a Ciência e a Tecnologia.

Data availability

The nucleotide sequences supporting this study are available from GenBank (accession numbers: KX681470 to KX681478; KX870949 to KX870952; KX870953

RESEARCH ARTICLE Journal of Experimental Biology (2016) 219, 3072-3081 doi:10.1242/jeb.140418

to KX870956; KX870957 to KX870960; KX870961 to KX870968; KX871034 to KX871041; KX871114 to KX871121; KX870969 to KX870978; KX870979 to KX870988; KX870989 to KX870999; KX871000 to KX871009; KX871010 to KX871015; KX871016 to KX871021; KX871022 to KX871027; KX871028 to KX871033; KX871042 to KX871053; KX871054 to KX871064; KX871065 to KX871077; KX871078 to KX871089; KX871090 to KX871095; KX871096 to KX871101; KX871102 to KX871107, KX871108 to KX871113; KX871122 to KX871132; KX871133 to KX871142; KX871143 to KX871150; KX871151 to KX871160; KX871161 to KX871166; KX871167 to KX871172; KX871173 to KX871178; KX871179 to KX871184).

Supplementary information

Supplementary information available online at
http://jeb.biologists.org/lookup/doi/10.1242/jeb.140418.supplemental

References

Adams, K. L. and Wendel, J. F. (2004). Exploring the genomic mysteries of polyploidy in cotton. *Biol. J. Linn. Soc.* **82**, 573-582.

Adams, K. L., Cronn, R., Percifield, R. and Wendel, J. F. (2003). Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proc. Natl. Acad. Sci. USA* **100**, 4649-4654.

Alves, M. J., Coelho, M. M. and Collares-Pereira, M. J. (2001). Evolution in action through hybridisation and polyploidy in an Iberian freshwater fish: a genetic review. *Genetica* **111**, 375-385.

Auger, D. L., Gray, A. D., Ream, T. S., Kato, A., Coe, E. H., Jr and Birchler, J. A. (2005). Nonadditive gene expression in diploid and triploid hybrids of maize. *Genetics* **169**, 389-397.

Blin, N. and Stafford, D. W. (1976). A general method for isolation of high molecular weight DNA from eukaryotes. *Nucleic Acids Res.* **3**, 2303-2308.

Bomblies, K. and Weigel, D. (2007). Hybrid necrosis: autoimmunity as a potential gene-flow barrier in plant species. *Nat. Rev. Genet.* **8**, 382-393.

Buggs, R. J. A., Zhang, L., Miles, N., Tate, J. A., Gao, L., Wei, W., Schnable, P. S., Barbazuk, W. B., Soltis, P. S. and Soltis, D. E. (2011). Transcriptomic shock generates evolutionary novelty in a newly formed, natural allopolyploid plant. *Curr. Biol.* **21**, 551-556.

Buggs, R. J., Wendel, J. F., Doyle, J. J., Soltis, D. E., Soltis, P. S. and Coate, J. E. (2014). The legacy of diploid progenitors in allopolyploid gene expression patterns. *Philos. Trans. R. Soc. B Biol. Sci.* **369**, pii:20130354.

Chelaifa, H., Monnier, A. and Ainouche, M. (2010). Transcriptomic changes following recent natural hybridization and allopolyploidy in the salt marsh species *Spartina×townsendii* and *Spartina anglica* (Poaceae). *New Phytol.* **186**, 161-174.

Chen, M. and Ni, M. (2006). RED AND FAR-RED INSENSITIVE 2, a RING-domain zinc finger protein, mediates phytochrome-controlled seedling deetiolation responses. *Plant Physiol.* **140**, 457-465.

Collares-Pereira, M. J., Matos, I., Morgado-Santos, M. and Coelho, M. M. (2013). Natural pathways towards polyploidy in animals: the *Squalius alburnoides* fish complex as a model system to study genome size and genome reorganization in polyploids. *Cytogenet. Genome Res.* **140**, 97-116.

Comai, L. (2000). Genetic and epigenetic interactions in allopolyploid plants. *Plant Mol. Biol.* **43**, 387-399.

Covelo-Soto, L., Leunda, P. M., Pérez-Figueroa, A. and Morán, P. (2015). Genome-wide methylation study of diploid and triploid brown trout (*Salmo trutta* L.). *Anim. Genet.* **46**, 280-288.

Cunha, C., Coelho, M. M., Carmona, J. A. and Doadrio, I. (2004). Phylogeographical insights into the origins of the *Squalius alburnoides* complex via multiple hybridization events. *Mol. Ecol.* **13**, 2807-2817.

Defossez, P.-A. and Stancheva, I. (2011). Biological functions of methyl-CpG-binding proteins. *Prog. Mol. Biol. Transl. Sci.* **101**, 377-398.

Diez, C. M., Roessler, K. and Gaut, B. S. (2014). Epigenetics and plant genome evolution. *Curr. Opin. Plant Biol.* **18**, 1-8.

Dreos, R., Ambrosini, G., Périer, R. and Bucher, P. (2015). The Eukaryotic Promoter Database: expansion of EPDnew and new promoter analysis tools. *Nucleic Acids Res.* **43**, D92-D96.

Farré, D., Bellora, N., Mularoni, L., Messeguer, X. and Albà, M. M. (2007). Housekeeping genes tend to show reduced upstream sequence conservation. *Genome Biol.* **8**, R140.

Feldman, M., Levy, A. A., Fahima, T. and Korol, A. (2012). Genomic asymmetry in allopolyploid plants: wheat as a model. *J. Exp. Bot.* **63**, 5045-5059.

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

Garcia, T. I., Matos, I., Shen, Y., Pabuwal, V., Coelho, M. M., Wakamatsu, Y., Schartl, M. and Walter, R. B. (2014). Novel method for analysis of allele specific expression in triploid *Oryzias latipes* reveals consistent pattern of allele exclusion. *PLoS ONE* **9**, e100250.

Grover, C. E., Gallagher, J. P., Szadkowski, E. P., Yoo, M. J., Flagel, L. E. and Wendel, J. F. (2012). Homoeolog expression bias and expression level dominance in allopolyploids. *New Phytol.* **196**, 966-971.

Hall, T. A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* **41**, 95-98.

Inácio, A., Pinho, J., Pereira, P. M., Comai, L. and Coelho, M. M. (2012). Global analysis of the small RNA transcriptome in different ploidies and genomic combinations of a vertebrate complex – the *Squalius alburnoides*. *PLoS ONE* **7**, e41158.

Ishikawa, R. and Kinoshita, T. (2009). Epigenetic programming: the challenge to species hybridization. *Mol. Plant* **2**, 589-599.

Kallman, K. D. (1975). The platyfish, *Xiphophorus maculates*. In *Handbook of Genetics* (ed. R. C. King), pp. 81-132. New York: Plenum Publishing.

Kerkel, K., Spadola, A., Yuan, E., Kosek, J., Jiang, L., Hod, E., Li, K., Murty, V. V., Schupf, N., Vilain, E. et al. (2008). Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. *Nat. Genet.* **40**, 904-908.

Kiryu, H., Oshima, T. and Asai, K. (2005). Extracting relations between promoter sequences and their strengths from microarray data. *Bioinformatics* **21**, 1062-1068.

Knudsen, S. (1999). Promoter 2.0: for the recognition of PolII promoter sequences. *Bioinformatics* **15**, 356-361.

Kuo, H. C., Lin, P. Y., Chung, T. C., Chao, C. M., Lai, L. C., Tsai, M. H. and Chuang, E. Y. (2011). DBCAT: database of CpG islands and analytical tools for identifying comprehensive methylation profiles in cancer cells. *J. Comput. Biol.* **18**, 1013-1017.

Lamatsch, D. K., Nanda, I., Epplen, J. T., Schmid, M. and Schartl, M. (2000). Unusual triploid males in a microchromosome-carrying clone of the Amazon molly, *Poecilia formosa*. *Cytogenet. Cell Genet.* **91**, 148-156.

Lamatsch, D. K., Stöck, M., Fuchs, R., Döbler, M., Wacker, R., Parzefall, J., Schlupp, I. and Schartl, M. (2010). Morphology, testes development and behaviour of unusual triploid males in microchromosome-carrying clones of *Poecilia formosa*. *J. Fish Biol.* **77**, 1459-1487.

Lamatsch, D. K., Trifonov, V., Schories, S., Epplen, J. T., Schmid, M. and Schartl, M. (2011). Isolation of a cancer-associated microchromosome in the sperm-dependent parthenogen *Poecilia formosa*. *Cytogenet. Genome Res.* **135**, 135-142.

Lampert, K. P. and Schartl, M. (2008). The origin and evolution of a unisexual hybrid: *Poecilia formosa*. *Philos. Trans. R Soc. B Biol. Sci.* **363**, 2901-2909.

Lampert, K. P., Lamatsch, D. K., Epplen, J. T. and Schartl, M. (2005). Evidence for a monophyletic origin of triploid clones of the Amazon molly, *Poecilia formosa*. *Evolution* **59**, 881-889.

Lampert, K. P., Lamatsch, D. K., Fischer, P., Epplen, J. T., Nanda, I., Schmid, M. and Schartl, M. (2007). Automictic reproduction in interspecific hybrids of poeciliid fish. *Curr. Biol.* **17**, 1948-1953.

Lee, T. Y., Chang, W. C., Hsu, J. B., Chang, T. H. and Shien, D. M. (2012). GPMiner: an integrated system for mining combinatorial cis-regulatory elements in mammalian gene group. *BMC Genomics* **13** Suppl 1, S3.

Li, J. and Zhang, Y. (2014). Relationship between promoter sequence and its strength in gene expression. *Eur. Phys. J. E Soft. Matter* **37**, 86.

Li, A., Hu, B.-Q., Xue, Z.-Y., Chen, L., Wang, W.-X., Song, W.-Q., Chen, C.-B. and Wang, C.-G. (2011). DNA methylation in genomes of several annual herbaceous and woody perennial plants of varying ploidy as detected by MSAP. *Plant Mol. Biol. Rep.* **29**, 784-793.

Mable, B. K. (2003). Breaking down taxonomic barriers in polyploidy research. *Trends Plant Sci.* **8**, 582-590.

Malone, C. D. and Hannon, G. J. (2009). Small RNAs as guardians of the genome. *Cell* **136**, 656-668.

Martienssen, R. A. and Colot, V. (2001). DNA methylation and epigenetic inheritance in plants and filamentous fungi. *Science* **293**, 1070-1074.

Matos, I., Sucena, E., Machado, M. P., Gardner, R., Inácio, A., Schartl, M. and Coelho, M. M. (2011). Ploidy mosaicism and allele-specific gene expression differences in the allopolyploid *Squalius alburnoides*. *BMC Genet.* **12**, 101.

Matzke, M. A., Scheid, O. M. and Matzke, A. J. M. (1999). Rapid structural and epigenetic changes in polyploid and aneuploid genomes. *Bioessays* **21**, 761-767.

McGaughey, D. M., Abaan, H. O., Miller, R. M., Kropp, P. A. and Brody, L. C. (2014). Genomics of CpG methylation in developing and developed zebrafish. *G3* **4**, 861-869.

Nanda, I., Schartl, M., Feichtinger, W., Schlupp, I., Parzefall, J. and Schmid, M. (1995). Chromosomal evidence for laboratory synthesis of a triploid hybrid between the gynogenetic teleost *Poecilia formosa* and its host species. *J. Fish Biol* **47**, 619-623.

Nanda, I., Schlupp, I., Lamatsch, D. K., Lampert, K. P., Schmid, M. and Schartl, M. (2007). Stable inheritance of host species-derived microchromosomes in the gynogenetic fish *Poecilia formosa*. *Genetics* **177**, 917-926.

Pala, I., Coelho, M. M. and Schartl, M. (2008). Dosage compensation by gene-copy silencing in a triploid hybrid fish. *Curr. Biol.* **18**, 1344-1348.

Pala, I., Schartl, M., Brito, M., Malta-Vacas, J. and Coelho, M. M. (2010). Gene expression regulation and lineage evolution: the North and South tale of the hybrid polyploid *Squalius alburnoides* complex. *Proc. R. Soc. B Biol. Sci.* **277**, 3519-3525.

Riddle, N. C. and Birchler, J. A. (2003). Effects of reunited diverged regulatory hierarchies in allopolyploids and species hybrids. *Trends Genet.* **19**, 597-600.

Journal of Experimental Biology (2016) 219, 3072-3081 doi:10.1242/jeb.140418

Salmon, A., Ainouche, M. L. and Wendel, J. F. (2005). Genetic and epigenetic consequences of recent hybridization and polyploidy in *Spartina* (Poaceae). *Mol. Ecol.* **14**, 1163-1175.

Schories, S., Lampert, K. P., Lamatsch, D. K., García de León, F. J. and Schartl, M. (2007). Analysis of a possible independent origin of triploid *P. formosa* outside of the Río Purificación river system. *Front. Zool.* **4**, 13.

Sehrish, T., Symonds, V. V., Soltis, D. E., Soltis, P. S. and Tate, J. A. (2014). Gene silencing via DNA methylation in naturally occurring *Tragopogon miscellus* (Asteraceae) allopolyploids. *BMC Genomics* **15**, 701.

Shen, H., He, H., Li, J., Chen, W., Wang, X., Guo, L., Peng, Z., He, G., Zhong, S., Qi, Y. et al. (2012a). Genome-wide analysis of DNA methylation and gene expression changes in two Arabidopsis ecotypes and their reciprocal hybrids. *Plant Cell* **24**, 875-892.

Shen, Y., Catchen, J., Garcia, T., Amores, A., Beldorth, I., Wagner, J., Zhang, Z., Postlethwait, J., Warren, W., Schartl, M. et al. (2012b). Identification of transcriptome SNPs between *Xiphophorus* lines and species for assessing allele specific gene expression within F_1 interspecies hybrids. *Comp. Biochem. Physiol. C Toxicol. Pharmacol.* **155**, 102-108.

Shen, Y., Garcia, T., Pabuwal, V., Boswell, M., Pasquali, A., Beldorth, I., Warren, W., Schartl, M., Cresko, W. A. and Walter, R. B. (2013). Alternative strategies for development of a reference transcriptome for quantification of allele specific expression in organisms having sparse genomic resources. *Comp. Biochem. Physiol. D Genomics Proteomics* **8**, 11-16.

Sousa-Santos, C., Robalo, J. I., Collares-Pereira, M.-J. and Almada, V. C. (2005). Heterozygous indels as useful tools in the reconstruction of DNA sequences and in the assessment of ploidy level and genomic constitution of hybrid organisms. *DNA Seq.* **16**, 462-467.

Sousa-Santos, C., Collares-Pereira, M. J. and Almada, V. C. (2006). Evidence of extensive mitochondrial introgression with nearly complete substitution of the typical *Squalius pyrenaicus*-like mtDNA of the *Squalius alburnoides* complex (Cyprinidae) in an independent Iberian drainage. *J. Fish Biol.* **68**, 292-301.

Sousa-Santos, C., Collares-Pereira, M. J. and Almada, V. (2007). Reading the history of a hybrid fish complex from its molecular record. *Mol. Phylogenet. Evol.* **45**, 981-996.

Stöck, M. and Lamatsch, D. K. (2013). Why comparing polyploidy research in animals and plants? *Cytogenet. Genome Res.* **140**, 75-78.

Stothard, P. (2000). The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *Biotechniques* **28**, 1102-1104.

Tirosh, I., Weinberger, A., Bezalel, D., Kaganovich, M. and Barkai, N. (2008). On the relation between promoter divergence and gene expression evolution. *Mol. Syst. Biol.* **4**, 159.

Turner, B. J., Balsano, J. S., Paul, J. M. and Rasch, E. M. (1983). Clonal diversity and evolutionary dynamics in a diploid-triploid breeding complex of unisexual fishes (*Poecilia*). *Evolution* **37**, 798-809.

Vinogradov, A. E. (1998). Genome size and GC-percent in vertebrates as determined by flow cytometry: the triangular relationship. *Cytometry* **31**, 100-109

Walia, H., Josefsson, C., Dilkes, B., Kirkbride, R., Harada, J. and Comai, L. (2009). Dosage-dependent deregulation of an AGAMOUS-LIKE gene cluster contributes to interspecific incompatibility. *Curr. Biol.* **19**, 1128-1132.

Wang, J., Tian, L., Lee, H.-S., Wei, N. E., Jiang, H., Watson, B., Madlung, A., Osborn, T. C., Doerge, R. W., Comai, L. et al. (2006). Genomewide nonadditive gene regulation in *Arabidopsis* allotetraploids. *Genetics* **172**, 507-517.

Wang, H., Jiang, J., Chen, S., Qi, X., Fang, W., Guan, Z., Teng, N., Liao, Y. and Chen, F. (2014). Rapid genetic and epigenetic alterations under intergeneric genomic shock in newly synthesized *Chrysanthemum morifolium* x *Leucanthemum paludosum* hybrids (Asteraceae). *Genome Biol. Evol.* **6**, 247-259.

Xiao, J., Song, C., Liu, S., Tao, M., Hu, J., Wang, J., Liu, W., Zeng, M. and Liu, Y. (2013). DNA methylation analysis of allotetraploid hybrids of red crucian carp (*Carassius auratus* red var.) and common carp (*Cyprinus carpio L.*). *PLoS ONE* **8**, e56409.

Yoo, M.-J., Szadkowski, E. and Wendel, J. F. (2013). Homoeolog expression bias and expression level dominance in allopolyploid cotton. *Heredity* **110**, 171-180.

Zhu, W.-G., Srinivasan, K., Dai, Z., Duan, W., Druhan, L. J., Ding, H., Yee, L., Villalona-Calero, M. A., Plass, C. and Otterson, G. A. (2003). Methylation of adjacent CpG sites affects Sp1/Sp3 binding and activity in the $p21^{Cip1}$ promoter. *Mol. Cell. Biol.* **23**, 4056-4065.

CHAPTER 5

Supplementary data

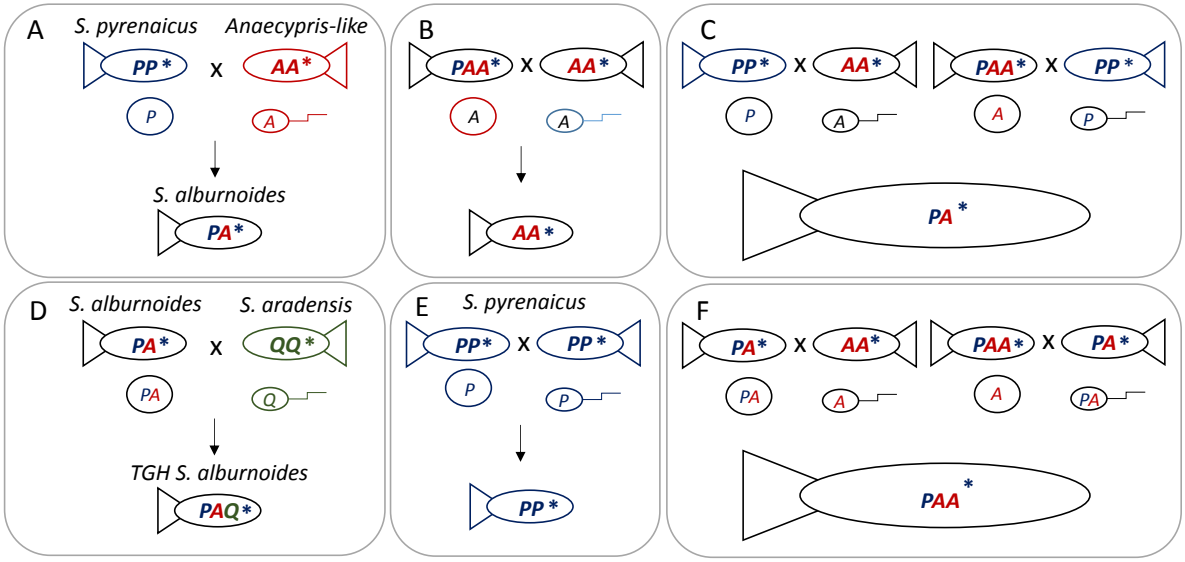


Fig. S1: Simplified overview of the *S. alburnoides* reproductive complex. A) Initial hybridization at the origin of the complex. B) *Anaecypris-like* nuclear genomotype reconstitution within the complex. C) Main crosses leading to diploid hybrid *S. alburnoides*. D) Cross leading to the artificial production of THG *S. alburnoides*. E) Independent maintenance of *S. pyrenaicus* species. F) Main crosses leading to triploid hybrid *S. alburnoides*. C and F exemplify the shift between ploidy levels and genomic compositions in each generation, and how the same genomotype can result from distinct crosses. Asterisk represents mitochondrial genotype: blue from *S. pyrenaicus,* red from *A. hispanica*-like and green from *S. aradensis*. This figure covers only the genomotypes involved in this study.

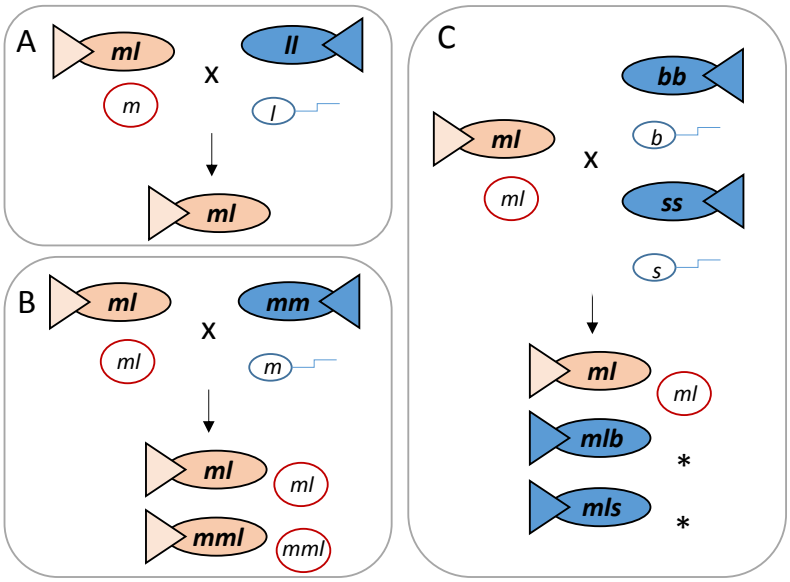


Fig. S2: Simplified overview of the *Poecilia formosa* complex. A) Initial hybridization in the origin of the complex. B) Gynogenetic diploid hybrid reproduction and paternal introgression leading to the occurrence of triploid hybrids. C) Crosses leading to the artificial production of THG *P. formosa*. *Pink fish are females, blue fish are males.* Asterisk represents sterility. This figure covers only the genomotypes involved in this study.

Table S1. List of specific Poecilia strains used in this work.

Species	Genomotype	Strain
Poecilia mexicana limantouri	*mm*	WLC1353
Poecilia latipinna	*ll*	WLC1368
Black molly	*bb*	WLC1351
Poecilia salvatoris	*ss*	WLC1330
Poecilia formosa 2n	*ml*	IV5
Poecilia formosa 3n	*mml*	WLC1055
Poecilia formosa 3n (TGH with b)	*mlb*	1588
		3331
		3830
		4069
		4335
		4664
		4665
Poecilia formosa 3n (TGH with s)	*mls*	1588
		1612
		3331
		4663
		4664

All strains are stocks in the aquarium of the Biocenter at the University of Würzburg, Germany.

Table S2. Primer sequences and references* for each target gene for both *Squalius* and *Poecilia.*

	Gene	Primer	Sequence	Reference
Squalius	***actb****	ß-ACTIN-F1	5'-CAACGGCTCCGGCATGTG-3'	Pala *et al.*, 2008
		ß-ACTIN-R1	5'-TGCCAGGGTACATGGTGG-3'	Pala *et al.*, 2008
	rpl8*	Rpl8 forward	5'-CTCCGTCTTCAAAGCCCATGT-3'	Pala *et al.*, 2008
		Rpl8 reverse	5'-TGTTCCTCGCAGTCTGCCAG-3'	Pala *et al.*, 2008
	Gapdh*	GAPDH-F1	5'-ATCAGGCATAATGGTTAAAGTTGG-3'	Pala *et al.*, 2008
		GAPDH-Ri	5'-GGCTGGGATAATGTTCTGAC-3'	Matos *et. al.,* 2010
Poecilia	***actb*****	ß-ACTIN-F1	5'-CAACGGCTCCGGCATGTG-3'	Pala *et al.*, 2008
		ß-ACTIN-R1	5'-TGCCAGGGTACATGGTGG-3'	Pala *et al.*, 2008
		Actin F pro	5'-CCTTAAAGCCCTGCCTACCT-3'	_
		Actin R pro	5'-AAGGGAAGGGATTGCTATGG-3'	_
	rpl8**	mRPL8F1	5´-ACGGAGTTTAGTGCACGAT-3'	_
		mRPL8R1	5'-CTTCTCCTGGACGGTCTTTG-3'	_
	rpl8**	Rpl8 F pro	5'-CTGTTTCCAYCCCCAGAAGT-3'	_
		Rpl8 R pro	5'-ACGATGCCCTTGATGTAGCC-3'	_
	Gapdh*	3gapdhF	5'-GTGACCCGWGCTGCTTTC-3'	_
		3gapdhR	5'-AGGTCACABACACGGTTGCT-3'	_
	Gapdh**	Gapdh F pro	5'-CATTTTGCRTTTTGTGGTTG-3'	_
		Gapdh R pro	5'-CCTCACATCKTGGTCTGAAA-3'	_

*PCR conditions as described in reference. **PCR conditions: pre-heating at 95°C for 3 min, 30 cycles at 95°C for 30 s, 60°C (*gapdh*)/58°C (*rpl8* and *ß-actin*)for 30 s and 72°C for 1 min and a final extension at 72°C for 10 min. ***PCR conditions: pre-heating at 95°C for 3 min, 30 cycles at 95°C for 30 s, 60°C (*gapdh*)/62°C (*rpl8* and *ß-actin*)for 30 s and 72°C for 1,30 min and a final extension at 72°C for 13 min.

Table S3. Relative comparison (ratio) between the mean observed methylation value and an expected methylation level in case of additivity (obs/ exp) for a hybrid situation.

		Liver			Muscle		
		Obs* 5mC (ng)	Exp**5mC (ng)	obs/exp	Obs* 5mC (ng)	Exp**5mC (ng)	obs/exp
S. alburnoides	***aa***	2,91	-	-	4,78	-	-
	pp	1,78	-	-	3,13	-	-
	qq	1,71	-	-	3,62	-	-
	pa	1,97	2,34	0,84	3,36	3,95	0,85
	paa	2,88	3,80	0,76	4,29	6,34	0,68
	pqa	3,17	3,20	0,99	4,69	5,76	0,81
P. formosa	***mm***	1,07	-	-	0,61	-	-
	ll	1,18	-	-	0,77	-	-
	bb	1,16	-	-	0,77	-	-
	ss	1,01	-	-	0,69	-	-
	ml	1,65	1,13	1,46	1,17	0,69	1,69
	mml	1,95	1,66	1,17	1,32	1,00	1,33
	mlb	0,98	1,71	0,58	0,81	1,08	0,76
	mls	1,03	1,63	0,63	0,78	1,04	0,75

* mean observed methylation value ** expected methylation calculated from the mean methylation level obtained for each parental diploid genomotype (*pp*, *aa* and *qq*) considering that *p*, *a* and *q* genomic contributions in the hybrids are methylated at the same level as in the non-hybrid situation.

Table S4. Identity matrix between all the parental genomes of *P. formosa* for each target gene from Bioedit.

	Seq *actb*->	*mm*	*ll*	*bb*	*ss*
actb	*mm*		0.983	0.991	0.976
	ll	0.983		0.978	0.977
	bb	0.991	0.978		0.974
	ss	0.976	0.977	0.974	

	Seq *gapdh*->	*mm*	*ll*	*bb*	*ss*
gapdh	*mm*		0.988	0.961	0.958
	ll	0.988		0.958	0.956
	bb	0.961	0.958		0.927
	ss	0.958	0.956	0.927	

	Seq *rpl8*->	*mm*	*ll*	*bb*	*ss*
rpl8	*mm*		0.97	0.987	0.98
	ll	0.97		0.967	0.966
	bb	0.987	0.967		0.972
	ss	0.98	0.966	0.972	

Table S5. Number of CpG sites per 1Kb within the *mm*, *ll*, *bb* and *ss* sequences, obtained with the Sequence Manipulation Suite.

		mm	*ll*	bb	ss
# CpG sites per 1Kb	***rpl8***	27	27	29	29
	gapdh	25	25	23	19
	actb	19	19	19	18

CHAPTER 6

Allele-specific expression variation at different ploidy levels in *Squalius alburnoides*

Matos I, Machado MP, Schartl M, Coelho MM. Allele-specific expression variation at different ploidy levels in Squalius alburnoides. Submitted for publication. (2018)

Allele-specific expression variation at different ploidy levels in *Squalius alburnoides*

Isa Matos* [1,2], Miguel P. Machado[1,2,3], Manfred Schartl[2,4,5], Maria Manuela Coelho[1]

[1], Faculdade de Ciências, cE3c- Centro de Ecologia, Evolução e Alterações Ambientais, Departamento de Biologia Animal, Universidade de Lisboa Campo Grande 1749-016 Lisboa, Portugal.

[2]University of Würzburg, Biozentrum, Physiological Chemistry, Am Hubland, Würzburg, Germany.

[3] Present Adress: Instituto de Microbiologia, Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Lisbon, Portugal.

[4]Comprehensive Cancer Center, University Clinic Würzburg, Josef Schneider Straße 6, 97074 Würzburg, Germany.

[5] Hagler Institute for Advanced Study and Department of Biology, Texas A&M University, College Station, USA

Abstract

Allopolyploid plants are long known to be subject to a homoeolog expression bias of varying degree. The same phenomenon was only much later suspected to occur also in animals based on studies of single selected genes in an allopolyploid vertebrate, the Iberian fish *Squalius alburnoides*. Consequently, this species became a good model for understanding the evolution of gene expression regulation in polyploid vertebrates. Here, we analyzed for the first-time genome-wide allele-specific expression data from diploid and triploid hybrids of *S. alburnoides* and compared homoeolog expression profiles of adult livers and of juveniles. Co-expression of alleles from both parental genomic types was observed for the majority of genes, but with marked homoeolog expression bias, suggesting homoeolog specific reshaping of expression level patterns in hybrids. Complete silencing of one allele was also observed irrespective of ploidy level, but not transcriptome wide as previously speculated. Instead, it was found only in a restricted number of genes, particularly ones with functions related to mitochondria and ribosomes. This leads us to hypothesize that allelic silencing may be a way to overcome intergenomic gene expression interaction conflicts, and that homoeolog expression bias may be an important mechanism in the achievement of sustainable genomic interactions, mandatory to the success of allopolyploid systems, as in *S. alburnoides*.

Introduction

By the classical Mendelian rules of inheritance for traits with intermediate phenotypes an equal contribution from the maternally and paternally inherited alleles to the overall expression was the intuitive solution. Doubts about this equal parental contribution to intermediate phenotypes have

been raised, especially in respect to hybrids and polyploids [1,2,3]. However, still to date, information's on allele-specific expression (ASE) are scarce and come just from a handful of organisms. To understand the biological meaning of ASE under physiological conditions and in organisms with special genomic situations like hybrids and polyploids, the high-throughput sequencing technologies allow to generate transcriptome-wide data from experimental model systems and non-model organisms.

In allopolyploid organisms, two or more sets of diverged genomes are joined through hybridization. Consequently, the gene copies that originated from each parent (homoeologs) may be quite different [4, 5]. The differences may be at the level of DNA sequence either in the promoter and/or within the transcribed region of a gene [4, 5]. Also, differences can be due to varied chromatin modifications and imprinting [6]. In any case, such differences between homoeologs may result in differential transcription rates and/or differential transcript decay between transcripts derived from each homoeolog, a phenomenon called homoeolog expression bias (HEB) [7, 8]. Studies on allopolyploids that have analyzed HEB are numerous for plants and go back several years [9,10, 11, 12, 13, 14, 15]. However, the same phenomenon was only much later described in a vertebrate organism [16, 17], the allopolyploid teleost fish *S. alburnoides* complex. Previously, based on the analysis of a set of 7 genes, it was shown that a gene-regulatory mechanism involving allelic silencing (AS), which is the most extreme case of HEB, contributes to the regulation of gene expression in allotriploid *S. alburnoides* individuals. The expression patterns were shown to vary according to the gene and organ analyzed, suggesting a considerable plasticity in the process and rejecting the hypothesis of whole haplome silencing [16, 17].

In addition to the phenomenon of AS, also gene expression dosage compensation was described to occur in *S. alburnoides*, reducing the allotriploid expression levels to the same levels of the hybrid diploid counterparts [16, 17, 18]. It was hypothesized that a link exists between AS and the observed dosage compensation [16]. It was also suggested that a consistent silencing of one of the three alleles (irrespective which one) across the allotriploid *S. alburnoides* genome could be a reason for the observed similar expression levels between diploid and triploid *S. alburnoides* specimens. To explore that hypotheses, a more detailed knowledge on how expression regulation occurs at the whole genome level in this allopolyploid species was necessary.

Following the initial discoveries of [15, 16], that were based just on a handful of genes, we applied for the first time in a naturally occurring allopolyploid vertebrate a whole transcriptome sequencing approach to the study of homoeolog specific expression (HSE) to help to clarify the role and implications of HEB, and AS, in the complex problem of odd genome regulation and allopolyploid perpetuation. We hoped to clarify if HEB and AS (balanced or unbalanced) are genome wide

transcriptomic phenomena or are restricted to subset of genes. Also important is to see if there is a preference towards one of the homoeologs to be down regulated or silenced and if there is a predisposition of certain genes to be affected by HEB. We hope also to clarify if and how does ploidy level increase affect HEB and AS.

In the present work we showed that HEB is quite extensive, but the full silencing of one of the alleles, which is the most extreme HEB scenario, is seen only for a minority of genes. Moreover, AS is not a genomewide transcriptomic phenomenon that systematically silences one of the alleles in *S. alburnoides* triploids, as previously thought.

Results

Genome specific expression patterns in livers of diploid and triploid hybrids.

We first considered SNVs for HSE quantification in liver tissue and used 2807 SNPs in liv-PA, distributed over 1121 transcripts, and 2305 SNPs in liv-PAA distributed over 937 transcripts. Plotting the distribution of transcripts according to the fraction of A genome contribution (Fig. 1) reveals the pattern of genome specific expression in diploid hybrids. Transcripts that exhibit a fraction larger than 0,5 are those that mapped more A than P reads in the hybrids. Those with ratios of less than 0,5 are transcripts from genes where the P alleles were expressed at a higher level than the A alleles. Following the criteria of [19], we assume as balanced allelic expression profile for diploids genes with less than 70% of expression preference of one the alleles (Fig. 1a). We found for liv-PA 764 of 1121 transcripts (68%) with balanced homoeolog expression and 357 transcripts (32%) showing strong homoeolog expression bias (Table 1). Additionally, we observed a significant unbalanced homoeolog expression bias (χ^2, $p<0,05$) towards P-genome alleles, meaning that for a significant number of genes displaying homoeolog expression bias, the P allele is preferentially expressed.

We then extrapolated the criteria used in [19] to triploids, and considered that for a ratio of two A alleles to one P allele, balanced expression between both genomotype alleles would result in an A genome expression fraction between 0,5 and 0,9 (Fig. 1b). For liv-PAA we found 649 of 938 transcripts (69%) with balanced homoeolog expression and 289 transcripts (31%) showing strong homoeolog expression bias (Table 1). Focusing on the transcripts presenting strong homoeolog expression bias, it was observed that as for the diploid form, the homoeolog expression bias is shifted significantly (χ^2, $P<0,05$) towards higher expression of the P alleles (Table 1).

From the transcripts that present strong allelic bias, we considered under allelic silencing the ones exhibiting a fraction of allelic contribution lower than 0,1. From the total analyzed genes 15% in

Liv-PA and 13% in Liv-PAA follow under this category (Table 1). Interestingly, we found that the number of transcripts presenting HEB is not significantly affected ($\chi^{2\text{- test}}$, P>0,05) by ploidy level, with both diploid and triploid showing similar number of genes with HEB, but also AS.

Genome specific contribution in hybrid diploid and triploid juveniles.

The effective number of SNVs considered for the HSE quantification was 5424 SNPs in juv-PA, distributed over 2039 transcripts and 5840 SNPs in juv-PAA distributed through 2169 transcripts. Again, we plotted the distribution of transcripts according to the fraction of A genome contribution, revealing the pattern of genome specific expression in triploid hybrids (Fig. 2). Using the same criteria as for the liver expressions, we found for juv-PA 1386 out of 2039 transcripts (68%) presenting balanced homoeolog expression and 653 transcripts (32%) showing strong homoeolog expression bias (Table 2). A significant unbalanced homoeolog expression bias (χ2-test, p<0,05) was also observed but different from liver. In Juv-PA the bias was towards A, meaning that from the pool of transcripts displaying HEB, for a significant majority, A alleles were higher expressed.

In juv-PAA we found 1064 SNPs in 2169 transcripts (49%) presenting balanced allelic expression and 1105 genes (51%) showing strong allele expression bias (Table 2). Significant unbalanced homoeolog expression bias ($\chi^{2\text{-test}}$; p<0,05) was observed towards P (Table 2). From the total analyzed genes, 8% in Juv-PA and 9% in Juv-PAA present (either A or P) monogenomic expression (Table 2).

In the case of the juvenile dataset, we found that, unlike in adult liver, the number of transcripts presenting HEB is significantly affected ($\chi^{2\text{- test}}$, P<0,05) by ploidy level.

Functional enrichment analysis.

We performed a GO and a KEGG pathway enrichment analysis in each of the defined monogenomic expression (MGE) groups of transcripts. We found significant functional enrichment of several terms (Tables 3 and 4). In essence, genes that underwent preferential silencing of the A alleles were enriched in ribosomal-linked terms (in juvenile samples, but also in both ploidy levels irrespective of the tissue type) while genes preferentially silenced for the P allele showed an enrichment of mitochondrial function related terms (in liver samples and both ploidy levels).

We further analised the genes affected by MGE and identified transcripts with consistent P or A monogenomic transcriptional contribution in all four libraries (Table 5). Within our criteria of significance, no functional enrichment was detected in either of these two groups. However, if considering the gene function of each of these above mentioned MGE genes, the link of allelic silencing

to mitochondria and to ribosomes is also seen.

Discussion:

In this work we describe homoeolog expression in the *S. alburnoides* complex, comparing the profiles of natural occurring diploids with triploids individuals to better understand the mechanisms of odd genome regulation and perpetuation in a successful allopolyploid vertebrate. It is the first transcriptomic attempt to quantify ASE in a natural allopolyploid fish.

Despite the PCR based approaches undertaken previously [16, 17] proved to be sensitive and valuable assays for an initial assessment of the gene expression profile in the *S. alburnoides* complex, technically they were constraint to the analysis of a few genes with a known sequence. Next-generation sequencing technologies brings together the advantages of high-throughput and high-sensitivity to the study of gene expression. The RNA-Seq approach allowed to look at the gene expression in the *S. alburnoides* complex at a much broader and integrative range. It allowed us to distinguish the different genome-specific gene copies and how they contribute to the overall expression of each gene for a much higher number of genes, and to draft a comparative profile of allele specific expression between diploid and triploid S*. alburnoides* fish.

When analyzing the genome-specific expression contribution per gene in the *S. alburnoides* complex, we found biased contribution of homoeologs that ranged from subtle differences to complete silencing of alleles, irrespective of the ploidy level or sample type. However, when comparing the results obtained for liver and for juveniles, results concordance is scarcer. We hypothesize that the lack of concordance between the liver and juveniles' datasets is mostly due to the intrinsic difference between a single organ and a whole animal. The impact of such differences on the output of gene expression profiling is well-documented [8, 20, 21, 22, 23, 24]. Thus, comparing homoeolog expression for such different sample types as liver and juvenile full body samples seems to be not very informative.

When analyzing the genome-specific expression contribution per gene in *S. alburnoides*, we found extreme biased contribution of homoeologs (as defined by [19]) in more than 30% of the considered transcripts, irrespective of the ploidy level or sample type. Hence, a considerably fraction of the genome is strongly affected by HEB. This result may even be an underestimation. As discussed by [25], the study of allele specific expression in full body samples and even in single tissue or organ samples can give skewed and/or diluted signals. The results we obtained imply that equally balanced

allelic expression is not a necessary regulatory condition to achieve appropriate amounts of gene product in fish of the *S. alburnoides* complex, neither as relevant factor to explain the success of allopolyploid *S. alburnoides* nor as distinctive mechanism between diploid and triploid *S. alburnoides* biotypes.

Gene expression is governed at the level of transcription by interactions between cis- and trans-acting regulatory elements [26, 27]. When in hybrids unequal expression of parental alleles has been observed, it has been considered as a signature of cis-regulatory divergence [26]. This may also apply for the significant degree of HEB which was found in the intergeneric hybrids of *S. alburnoides*. Evidence is accumulating that also in non-hybrid genomes, the variation found in regulatory regions between alleles is sufficient to affect the level of expression of the two variants [4, 23]. For example, significant cis regulatory variation in 80% of mouse genes have been found [28] and allelic imbalance was estimated to affect greater than 89% of genes of mouse, cow and humans, in at least one tissue [23].

A different picture emerges from a study of another hybrid fish, the Amazon molly (*P. formosa*). Here, allele specific gene expression analysis from different organs, including liver, brain and ovary revealed only a very small percentage of genes (between 1.2 and 4.1%) presenting HEB [29]. However, *P. formosa* is a clonal hybrid organism, resultant from a *single time* successful hybridization event at least 100.000 years old [29, 30]. This makes *P. formosa* much different from the reproductive complex of S*. alburnoides,* which results from a continuum of intricated networks of genetic exchanges, *de novo* hybridizations and ploidy levels shifts [31]. The old "frozen" hybrid genomic context of *P. formosa* may have evolved mechanism that counteract HEB.

An important feature of the homoeolog expression profile is whether this expression is balanced or unbalanced [7]. Balanced homoeolog expression means that expression does not favor one component genome, while in unbalanced homoeolog expression one of the intervenient genomes is favored [8]. We found a significant unbalanced homoeolog expression in all *S. alburnoides* libraries. In all liver samples there was consistency in terms of magnitude and bias direction towards the P genome. Conversely, in the whole-body juvenile samples, the ploidy state appears to influence the direction of the bias. HEB was skewed towards A in Juv-PA sample while it is displaced towards P in juv-PAA. In liver, ploidy level did not significantly affect neither the extent of HEB nor the tendency towards preferential expression of the P alleles. Those results imply that balanced allelic expression is not a regulatory necessity to cope with elevated ploidy in *S. alburnoides.* Our data also supports the notion that homoeolog specific expression in diploid and triploid *S. alburnoides* liver is not a simple additive phenomenon. It was previously shown [18] that the quantitative expression profiles of livers

from the *S. alburnoides* parental genomotypes (AA and PP) are significantly different. Most transcripts were found at much higher levels in AA than in PP livers [18]. Taken this into consideration, a simplistic model of additive homoeolog expression in allopolyploid *S. alburnoides* could be only put forward if we had found homoeolog expression bias towards A homoeologs, but not towards P, as it was the case. As many interactions between divergent regulatory machineries occur, new patterns of gene expression and homoeolog regulation may be more complex and difficult to predict. In plants, reshaping of homoeolog expression has been commonly found [5, 26, 32]. It was also noted in this context that alterations in to the original expression pattern of the originally non-dominant genome occurred [32].

Our results from *S. alburnoides*, are in line with studies from other systems, where unbalanced expression has been commonly observed in plant hybrids of different ploidies [11, 12, 33, 34, 35, 36, 37]. Notably, in cotton, significant differences between studies, in terms of the magnitude of the expression bias and bias direction, have been found [32, 38, 39].

While we observed for liver samples that the extent of HEB was not significantly affected by ploidy level (2n vs 3n), in juveniles there was a significantly higher number of transcripts in 3n than in the 2n juveniles presenting HEB.

Unequal expression of parental alleles has been pointed out in diploid hybrid plants as a signature of cis-regulatory divergence, because both parental alleles should be proportionally exposed to the same set of trans-acting regulators [26]. However, for an unorthoploid (increased and uneven ploidy level hybrid) this assumption is not straightforward since the network of interactions between cis and trans regulators is unpredictably influenced by the unbalanced contribution of parental genomes and increased number of non-additive interactions between the parental genomes [5].

The extent of HEB ranged in our analysis from only subtle allelic differences to complete silencing of one (or more) alleles. Thus, we considered another phenomenon, allele-specific silencing or monogenomic gene expression, where expression is derived from only one of the parental genomes In diploid hybrids, when transcription from only one allele was detected it is undoubtful to infer that the other allele is silenced. In the case of allotriploid *S. alburnoides* of PAA genomic composition, when only P genome derived expression was detected at any locus this has to result from silencing of both A homoeologs at that locus. However, in the case of expression only from the A genome we cannot conclude if A transcripts are coming from one or from both A alleles. Thus, exclusive expression of A can mean either biallelic or monoallelic expression.

In diploid *S. alburnoides* (PA), even though most transcripts analyzed presented biallelic

expression, we detected for the first time in this fish model the occurrence of MGE, either from the A or P homoeologs. This is in accordance with several studies on other diploid organisms, where transcription from only one allele has been found not only due to sex-chromosome inactivation and genomic imprinting but also stochastic silencing in autosomal genes (reviewed in [25]).

In triploid *S. alburnoides* (liv-PAA and juv-PAA) we found, besides the biallelic expression, monoallelic expression of P but also P allelic silencing. P homoeolog silencing in triploid *S. alburnoides* has already been reported previously [16, 17, 18, 31], but monoallelic P expression has not been observed so far in *Squalius* genus. This new finding of expression from only one allele in the context of allotriploidy in this fish, agrees with a previous report of the same pattern in another successful allopolyploid complex, the allopolyploid *Poecilia formosa* [31].

To explain the first observations of P allele specific silencing in triploid *S. alburnoides*, based on the analysis of a limited number of genes [16], a parsimonious hypothesis was suggested postulating that one of the three alleles, irrespective which one, could be systematically silenced across the entire *S. alburnoides* genome. This was proposed as explanation why triploids presented similar expression levels to their diploid counterparts for the set of analyzed genes. However, our genome wide analyses show that allelic silencing does not happen genome-wide in triploid *S. alburnoides,* and additionally shows that AS can also be found in diploid S. alburnoides, and at the same extension than in the triploids.

Despite heterosis and hybrid vigor are well known phenomena associated with hybrids [40], not all crosses result in heterosis and some hybrids do not even survive and/or reproduce [41, 42]. Traits derived from different genetic backgrounds merged in the hybrids may not be fully compatible, and fitness can be reduced. A possible explanation for the success of some hybrids like *S. alburnoides* may come from gene expression plasticity [18], where ASE regulation at each locus may have a significant role.

We investigated also the biological context of the AS occurrence in *S. alburnoides*. As mentioned, upon hybridization (and ploidy increase) disruption of well-established interlocus interactions may reveal incompatibilities [41, 42], so many hybrids and allopolyploids may either be non-viable or suffer from reduced fitness. Mito-nuclear incompatibilities have been found to influence hybrid inviability [43, 44]. More specifically, there is evidence of hybrid incompatibilities between nuclear- and mitochondrial DNA (mtDNA)-encoded elements [45], for example the interaction between nuclear- and mtDNA-encoded subunits of the OXPHOS proteins [46, 47].

In the *S. alburnoides* complex, apart from a few exceptions [48, 49], there is almost exclusive

presence of *S. pyrenaicus* (P) mtDNA [50]. In that sense, it is interesting to note that P monogenomic expression was associated to mitochondria related GO terms, irrespective of ploidy level and sample type. We thus hypothesise that expression of only the P alleles of mitochondria related loci in *S. alburnoides* specimens might be an effective way to cope with incompatibilities of the hybrid genome and the P derive mitochondria [45]. For instance, by facilitating or optimizing mitochondrial-nuclear interactions through reducing post-transcriptional and translational incompatibilities between the PA(A) nuclear DNA and the maternally inherited P-only mtDNA. This is in accordance with the mitonuclear coadaptation theory [51], which postulates that nuclear genes that interact with mitochondria are expected to be maternally biased. Also, we found a tendency towards monogenomic A transcriptional activity of genes related to ribosomes. Assembly of ribosomes involves more than 300 proteins and RNAs [52]. The genes that code for RNA molecules constitute the ribosomal sub-units and are organized in tandem repeats at chromosomal regions called Nucleolar Organizing Regions (NORs) [53]. But, not all NORs are transcriptionally active. It was previously found that *S. pyrenaicus* presented only one pair of chromosomes with active NORs, while all forms of *S. alburnoides* presented mostly multichromosomal active NORs. Hence, the observed increased NOR numbers in the *S. alburnoides* complex specimens would be A genome derived [54]. Accordingly, it can be assumed that at a given time, there is a high probability to find more A-genome derived than P-genome derived ribosomal RNA molecules, as both PA and PAA genomotype individuals would have only one P-derived competent NOR per cell while having multiple A-derived NORs. Also, in accordance with the gene balance hypothesis [55,56] which posits that in multi-subunit complexes, changes in the stoichiometry of the components of those complexes is deleterious, an intergenomic ribosomal gene conflict can be speculated to support the tendency we found towards monogenomic A transcriptional activity of ribosomal related genes. Functional studies are required to substantiate these considerations.

In conclusion, our results imply that balanced allelic expression is not a necessary regulatory condition to achieve appropriate amounts of gene products in the *S. alburnoides* complex and support that homoeolog specific expression in diploid and triploid *S. alburnoides* is not a simple additive phenomenon.

Despite HEB is quite extensive, the full silencing of one of the alleles as the extreme, was seen only in a minority of genes. However, AS was found mostly in genes related to mitochondria and ribosomes, what lead us to hypothesize that AS may be a way to overcome intergenomic gene expression interaction conflicts. In that sense, HEB and AS may be key players into the achievement

of sustainable genomic interactions, mandatory to the success of allopolyploid systems, as the *S. alburnoides* complex.

Materials and methods

Model system

We used the allopolyploid hybridogenetic complex *S. alburnoides* as experimental model to study specific allelic contribution to the transcript pool. "Hybridogenetic" refers to an alternative mode of reproduction and "complex" is the terminus denoting a natural system composed of parental species and their hybrids of different ploidies, with altered modes of reproduction and reproductive interdependence. The *S. alburnoides* complex resulted from a cross of a *Squalius pyrenaicus* female (contributing with the P genome) and an *Anaecypris-like* male (contributing with the A genome) (see [50] and [57] for extensive review).

Libraries

Previously constructed and sequenced RNA-Seq libraries, enriched for mRNA by hybridization with Oligo-dT beeds, have been used in this study (Supplementary Table S1 – all additional files at https://figshare.com/s/c03974866dbd92b5a24d).

In summary:

- From juvenile samples: Three barcoded RNA libraries had been previously constructed and paired-end sequenced using Illumina HiSeq 2000, producing 12 Gb clean data in 3 data sets (juv-AA; juv-PA; and juv-PAA, ~4Gb per library) of paired-end sequence reads (around 91 bp) [18].
- From adult liver samples: Four barcoded RNA libraries had been previously constructed, one for *S. pyrenaicus* (liver- PP) and three for *S. alburnoides* (liver- AA, liver-PA and liver-PAA) [18]. The four libraries were paired-end sequenced using Illumina HiSeq 2000, producing 4Gb of clean data in 4 data sets (Liv-AA; Liv-PA; Liv-PP and Liv-PAA, ~1Gb per library) of short paired-end sequence reads (around 50 bp) [18].
- From adult brain and gonad samples: Six barcoded RNA libraries were constructed, two for *S. pyrenaicus* brain and two for gonads [58], one for *S. alburnoides* nuclear non-hybrid AA male brain and one for its gonad (this study). For the construction of the six libraries, total RNA was purified from individual gonads and brains of two *S. pyrenaicus* (PP) individuals, and from a

pool of brains and gonads (separated by tissue type) of a AA *S. alburnoides* and a rare occurring AAA *S. alburnoides.* All libraries were paired-end sequenced using Illumina HiSeq 2000, producing in total 12 Gb of clean data in six data sets (brainF-PP, brainM-PP, gonF-PP, gonM-PP; brain-AA and gon-AA, ~2 Gb per library) of paired-end sequence reads (around 91 bp).

The RNA-Seq fastq files are available through public repositories (Supplementary Table S1 - Additional Files at https://figshare.com/s/c03974866dbd92b5a24d).

S. alburnoides "de novo" transcriptome assembly

We used SOAPdenovo to produce a transcriptome *de novo* assembly using all available libraries from individuals participating in the *S. alburnoides* complex (juv-AA, juv-PA, juv-PAA, liv-PP, liv-AA, liv-PA, liv-PAA, brainF-PP, brainM-PP, gonF-PP, gonM-PP, brain-AA and gon-AA) to produce a more comprehensive "*S. alburnoides* breeding complex" reference transcriptome than the one previously available at [18]. Statistics of assembly quality for *S. alburnoides* complex transcriptome provided as Supplementary Table S2 (all additional files at https://figshare.com/s/c03974866dbd92b5a24d). Assemblies were taken into further processes of sequence splicing and redundancy removing with the sequence clustering software TGICL [59]. After clustering, contigs were annotated with blastx and blastn against the NCBI non-redundant protein database (NR) (e-value<0.00001), retrieving proteins with the highest sequence similarity to the given contigs. We used Blast2GO program [60] to get functional annotations. Assembly and gene ontology (GO) annotations are available as Datasets in Additional Files at https://figshare.com/s/c03974866dbd92b5a24d.

Using the *de novo* assembled *S. alburnoides* complex transcriptome as reference, we detected and quantified single nucleotide variants (SNVs) between all different samples with SOAPsnp [61]. SNV calling criteria was as following: consensus quality ≥ 20; depth of coverage of the site and the flanking sequences ≥ 3; distance from the last candidate SNV ≥ 5bp; distance from the borders > 5bp. SNP calling, and alleles quantification are provided as Supplementary Information.

Comparative genome specific expression quantification

Several polymorphic sites were detected within each nuclear non-hybrid sample (both *S. pyrenaicus* and *S. alburnoides* AA samples) and binned for separation from the non-polymorphic sites. After excluding the intragenomic polymorphisms, P and A genome specific variants were identified. Only when a single nucleotide variant was detected within all 4 AA genome libraries, and it was different from the single variant found in all 5 PP genome libraries, such site was considered for the

allele specific expression quantification in the hybrid *S. alburnoides* individuals. In our reference transcriptome sequences, which represents a hybrid species, the polymorphic positions are represented by the most frequent or more represented nucleotide base from the pool of all reads covering that site. Thus, as observed by [19], reads of the same SNV as reference, map with higher efficiency than others, an effect known as reference bias. When inspecting our data, it was obvious that there was a clear tendency towards higher read counts of the reference variant. To validate and calculate this read count bias, we used the intra-genomic polymorphisms. To do so, we started with the assumption of a 1:1 allelic contribution in the complex nuclear non-hybrid forms (*S. pyrenaicus* and *S. alburnoides* AA samples) and calculated the deviation from the expectation (allele1 reads / allele2 reads = 1) at each position. Mean deviation was calculated for each nuclear non-hybrid sample, and a mean value of these means was used as correction factor of read counts of each transcript of the hybrid libraries. brain-AA and gon-AA were excluded from this analysis because of the inclusion of triploid AAA in the AA sample pools.

In the hybrid samples, only nucleotide positions showing 20 or more SNV supporting reads were considered for quantification, increasing the confidence of the quantitative allele/genome-specific expression analysis, but obviously reducing the number of SNV positions to analyze. For the majority of transcripts identified as having SNVs between P and A genomes, more than one site per transcript was identified. In these cases, for each hybrid dataset a mean number of reads per allele/genome and per transcript was calculated.

The number of reads for each variant of the polymorphic site per transcript in the hybrid libraries (PA and PAA) was used to quantify the genome specific and/or allele specific contribution for the overall expression.

Functional term enrichment analysis

To infer a possible biological context of genes exhibiting preferential AS in *S. alburnoides*, transcripts in which expression was determined to be coming only from one genome type (P or A) were organized in groups according to genome-specific silencing within tissue/sample type (liver or juveniles) and also according to ploidy level (2n-PA and 3n-PAA). 8 groups of monogenomic derived transcripts were assembled as follows: within liver libraries (independently of ploidy level), group i) of transcripts presenting P monogenomic expression and group ii) presenting A monogenomic expression; within juveniles libraries (independently of ploidy level), group iii) of transcripts presenting P monogenomic expression and group iv) presenting A monogenomic expression; within the PA

libraries (independently of sample type), group v) of transcripts presenting P monogenomic expression and group vi) presenting A monogenomic expression; within the PAA libraries (independently of sample type), group vii) of transcripts presenting P monogenomic expression and the group viii) presenting A monogenomic expression. The list of transcripts organized according to genome-specific silencing (from i to viii) is available as Supporting Information in Additional Files (https://figshare.com/s/c03974866dbd92b5a24d).

To perform functional enrichment analysis in each of these groups, we used DAVID Bioinformatics Resource v6.7 (http://david.abcc.ncifcrf.gov/), with default parameters. The top blastx hits in nr database corresponding to each *S. alburnoides* contigs were used as customized reference background and compared to the above-mentioned input lists of monogenomic transcribed genes.

Enriched terms were ranked in the ontology categories Biological Process (BP), Cellular Component (CC) and Molecular Function (MF) and KEGG pathways.

Significant enrichment was only considered when Benjamini corrected p-value was ≤ 0.05.

References:

1. Gimelbrant, A., Hutchinson, J. N., Thompson, B. R. & Chess, A. Widespread monoallelic expression on human autosomes. *Science* **318**, 1136-40 (2007). 10.1126/science.1148910

2. Prestel, M., Feller, C. & Becker, P. B. Dosage compensation and the global re-balancing of aneuploid genomes. *Genome Biology* **11,** 216 (2010). http://doi.org/10.1186/gb-2010-11-8-216

3. Hegarty, M. Hybridization: expressing yourself in a crowd. *Current Biology* **21**, R254-5 (2011). https://doi.org/10.1016/j.cub.2011.02.035

4. McManus, C. J. *et al.* Regulatory divergence in *Drosophila* revealed by mRNA-seq. *Genome Research* **20**, 816–825 (2010). http://doi.org/10.1101/gr.102491.109

5. Combes, M.-C. *et al.* Regulatory Divergence between Parental Alleles Determines Gene Expression Patterns in Hybrids. *Genome Biology and Evolution* **7**, 1110–1121 (2015). http://doi.org/10.1093/gbe/evv057

6. Knight, J. C. Allele-specific gene expression uncovered. *Trends in Genetics* **20**, 113-116 (2004). https://doi.org/10.1016/j.tig.2004.01.001

7. Chen, Z. J. Genetic and Epigenetic Mechanisms for Gene Expression and Phenotypic Variation in Plant Polyploids. *Annual Review of Plant Biology* **58**, 377-406 (2007). http://doi.org/10.1146/annurev.arplant.58.032806.103835

8. Grover, C. E. *et al*. Homoeolog expression bias and expression level dominance in allopolyploids. *New Phytologist* **196**, 966-971 (2012). https://doi.org/10.1111/j.1469-8137.2012.04365.x

9. Guo, M., Davis, D. & Birchler, J. A. Dosage Effects on Gene Expression in a Maize Ploidy Series. *Genetics* **142**, 1349–1355 (1996).

10. Auger, D.L. *et al*. Nonadditive gene expression in diploid and triploid hybrids of maize. *Genetics* **169**, 389-97 (2005). https://doi.org/10.1534/genetics.104.032987

11. Flagel, L., Udall, J., Nettleton, D. & Wendel, J. Duplicate gene expression in allopolyploid *Gossypium* reveals two temporally distinct phases of expression evolution. *BMC Biology* **6**, 16 (2008). http://doi.org/10.1186/1741-7007-6-16

12. Chaudhary, B. *et al.* Reciprocal Silencing, Transcriptional Bias and Functional Divergence of Homeologs in Polyploid Cotton (Gossypium). *Genetics* **182**, 503–517 (2009). http://doi.org/10.1534/genetics.109.102608

13. Buggs, R. J. *et al*. Transcriptomic shock generates evolutionary novelty in a newly formed, natural allopolyploid plant. *Current Biology* **21**, 551-556 (2011). https://doi.org/10.1016/j.cub.2011.02.016

14. Dong, S. & Adams, K. L. Differential contributions to the transcriptome of duplicated genes in response to abiotic stresses in natural and synthetic polyploids. *New Phytologist* **190**, 1045-1057 (2011). https://doi.org/10.1111/j.1469-8137.2011.03650.x

15. Combes, M. C., Cenci, A., Baraille, H., Bertrand, B. & Lashermes, P. Homeologous gene expression in response to growing temperature in a recent allopolyploid (Coffea arabica L.). *Journal of Heredity* **103,** 36–46 (2012). https://doi.org/10.1093/jhered/esr120

16. Pala, I., Coelho, M. M. & Schartl, M. Dosage compensation by gene-copy silencing in a triploid hybrid fish. *Current Biology* **18**, 1344-1348 (2008). https://doi.org/10.1016/j.cub.2008.07.096

17. Pala, I., Schartl, M., Brito, M., Vacas, J. M. & Coelho, M. M. Gene expression regulation and lineage evolution: the North and South tale of the hybrid polyploid *Squalius alburnoides* complex. *Proceedings of the Royal Society B: Biological Sciences **277***, 3519–3525 (2010). http://doi.org/10.1098/rspb.2010.1071

18. Matos, I., Machado, M. P., Schartl, M. & Coelho, M. M. GeneExpression Dosage Regulation in an Allopolyploid Fish. *PLoS ONE* **10**, e0116309 (2015). http://doi.org/10.1371/journal.pone.0116309

19. Shen, Y. *et al.* Identification of transcriptome SNPs between *Xiphophorus* lines and species for assessing allele specific gene expression within F1 interspecies hybrids. *Comparative Biochemistry and Physiology. Toxicology & Pharmacology CBP* **155**, 102–108 (2012). http://doi.org/10.1016/j.cbpc.2011.03.012

20. Adams, K. L., Cronn, R., Percifield, R. & Wendel, J. F. Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 4649–4654 (2003). http://doi.org/10.1073/pnas.0630618100

21. Adams, K. L., Percifield, R. & Wendel, J. F. Organ-Specific Silencing of Duplicated Genes in a Newly Synthesized Cotton Allotetraploid. *Genetics* **168**, 2217–2226 (2004). http://doi.org/10.1534/genetics.104.033522

22. Whitehead, A. & Crawford, D. L. Variation in tissue-specific gene expression among natural populations. *Genome Biology* **6**, R13 (2005). http://doi.org/10.1186/gb-2005-6-2-r13

23. Chamberlain, A. J. *et al*. Extensive variation between tissues in allele specific expression in an outbred mammal. *BMC Genomics* **16**, 993 (2015). http://doi.org/10.1186/s12864-015-2174-0

24. Pinter, S. F. *et al.* Allelic Imbalance Is a Prevalent and Tissue-Specific Feature of the Mouse Transcriptome. *Genetics* **200**, 537-549 (2015). http://doi.org/10.1534/genetics.115.176263

25. Eckersley-Maslin, M. A. & Spector, D. L. Random Monoallelic Expression: Regulating gene expression one allele at a time. *Trends in Genetics: TIG* **30**, 237-244 (2014). http://doi.org/10.1016/j.tig.2014.03.003

26. Bell, G. D. M., Kane, N. C., Rieseberg, L. H. & Adams, K. L. RNA-Seq Analysis of Allele-Specific Expression, Hybrid Effects, and Regulatory Divergence in Hybrids Compared with Their Parents from Natural Populations. *Genome Biology and Evolution* **5**, 1309–1323 (2013). http://doi.org/10.1093/gbe/evt072

27. Xu, C., *et al.* Genome-Wide Disruption of Gene Expression in Allopolyploids but Not Hybrids of Rice Subspecies. *Molecular Biology and Evolution*, **31**, 1066–1076 (2014). http://doi.org/10.1093/molbev/msu085

28. Crowley, J. J., *et al.* Analyses of Allele-Specific Gene Expression in Highly Divergent Mouse Crosses Identifies Pervasive Allelic Imbalance. *Nature Genetics* **47**, 353–360 (2015). http://doi.org/10.1038/ng.3222

29. Warren, W. C. *et al*. Clonal polymorphism and high heterozygosity in the celibate genome of the Amazon molly. *Nature Ecology & Evolution* **2**, 669-679 (2018). 10.1038/s41559-018-0473-y

30. Stöck, M., Lampert, K.P., Möller, D., Schlupp, I. & Schartl, M. Monophyletic origin of multiple clonal lineages in an asexual fish (Poecilia formosa). *Molecular Ecology* **19**, 5204-15 (2010). https://doi.org/10.1111/j.1365-294X.2010.04869.x

31. Matos, I.M., Coelho, M. M. & Schartl, M. Gene copy silencing and DNA methylation in natural and artificially produced allopolyploid fish. *Journal of Experimental Biology* **219,** 3072-3081 (2016). 10.1242/jeb.140418

32. Yoo, M.-J., Szadkowski, E. & Wendel, J. F. Homoeolog expression bias and expression level dominance in allopolyploid cotton. *Heredity* **110**, 171–180 (2013). http://doi.org/10.1038/hdy.2012.94

33. Chen, Z. J. & Pikaard, C. S. Transcriptional analysis of nucleolar dominance in polyploid plants: Biased expression/silencing of progenitor rRNA genes is developmentally regulated in *Brassica*. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 3442–3447 (1997).

34. Wang, J. *et al.* Genomewide Nonadditive Gene Regulation in Arabidopsis Allotetraploids. *Genetics* **172**, 507–517 (2006). http://doi.org/10.1534/genetics.105.047894

35. Akhunova, A. R., Matniyazov, R. T., Liang, H. & Akhunov, E. D. Homoeolog-specific transcriptional bias in allopolyploid wheat. *BMC Genomics* **11**, 505 (2010). http://doi.org/10.1186/1471-2164-11-505

36. Buggs, R. J. *et al.* Tissue-specific silencing of homoeologs in natural populations of the recent allopolyploid Tragopogon mirus. *New Phytologist Trust* **186**, 175-83 (2010). https://nph.onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-8137.2010.03205.x

37. Schnable, J. C. & Freeling, M. Genes Identified by Visible Mutant Phenotypes Show Increased Bias toward One of Two Subgenomes of Maize. *PLoS ONE* **6,** e17855 (2011). http://doi.org/10.1371/journal.pone.0017855

38. Rapp, R. A., Udall, J. A. & Wendel, J. F. Genomic expression dominance in allopolyploids. *BMC Biology* **7**, 18 (2009). http://doi.org/10.1186/1741-7007-7-18

39. Flagel, L. E. & Wendel, J. F. Evolutionary rate variation, genomic dominance and duplicate gene expression evolution during allotetraploid cotton speciation. *New Phytologist* **186**, 184–193 (2010). https://doi.org/10.1111/j.1469-8137.2009.03107.x

40. Baranwal, V. K., Mikkilineni, V., Zehr, U. B., Tyagi, A. K., & Kapoor, S. Heterosis: emerging ideas about hybrid vigor. *Journal of Experimental Botany* **63**, 6309-14 (2012). https://doi.org/10.1093/jxb/ers291

41. Walia, H., Wilson, C., Ismail, A. M., Close, T. J. & Cui, X. Comparing genomic expression patterns across plant species reveals highly diverged transcriptional dynamics in response to salt stress. *BMC Genomics* **10**, 398 (2009). http://doi.org/10.1186/1471-2164-10-398

42. Malone, C. D. & Hannon, G. J. Small RNAs as Guardians of the Genome. *Cell* **136**, 656–668 (2009). http://doi.org/10.1016/j.cell.2009.01.045

43. Trier, C. N., Hermansen, J. S., Sætre, G.-P. & Bailey, R. I. Evidence for Mito-Nuclear and Sex-Linked Reproductive Barriers between the Hybrid Italian Sparrow and Its Parent Species. PLoS Genetics **10**, e1004075 (2014). http://doi.org/10.1371/journal.pgen.1004075

44. Bundus, J. D., Wang, D. & Cutter, A. D. Genetic basis to hybrid inviability is more complex than hybrid male sterility in Caenorhabditis nematodes. *Heredity* **121**, 169-182 (2018). doi: 10.1038/s41437-018-0069-y.

45. Lane, N. Mitonuclear match: optimizing fitness and fertility over generations drives ageing within generations. *Bioessays* **33**, 860-869 (2011). https://doi.org/10.1002/bies.201100051

46. Burton, R. S., Ellison, C. K. & Harrison, J. S. The sorry state of F2 hybrids: consequences of rapid mitochondrial DNA evolution in allopatric populations. *The American Naturalist* **168**, S14-S24 (2006). https://www.journals.uchicago.edu/doi/10.1086/509046

47. Ellison, C. K. & Burton, R. S. Disruption of mitochondrial function in interpopulation hybrids of Tigriopus californicus. *Evolution* **60**, 1382-1391 (2006).

48. Alves, M.J., Coelho, M.M., Collares-Pereira, M.J. & Dowling, T.E. Maternal ancestry of the Rutilus alburnoides comples (Teleostei, Cyprinidae) as determined by analysis of cytochrome b sequenbces. *Evolution* **51**, 1584-1592 (1997). https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1558-5646.1997.tb01481.x

49. Sousa-Santos, C., Collares-Pereira, M. J. & Almada, V. C. Evidence of extensive mitochondrial introgression with nearly complete substitution of the typical Squalius pyrenaicus-like mtDNA of the Squalius alburnoides complex (Cyprinidae) in an independent Iberian drainage. *Journal of fish biology* **68**, 292-301 (2006). https://doi.org/10.1111/j.0022-1112.2006.01081.x

50. Alves, M.J., Coelho, M.M. & Collares-Pereira, M. J. Evolution in action through hybridisation and polyploidy in an Iberian freshwater fish: a genetic review. *Genetica* **111**, 375-85 (2001).

51. Wolf, J. B. Cytonuclear interactions can favor the evolution of genomic imprinting. *Evolution* **63**, 1364–1371 (2009).

52. Staley, J. P. & Woolford, J. L. Assembly of ribosomes and spliceosomes: complex ribonucleoprotein machines. *Current Opinion in Cell Biology* **21**, 109–118 (2009). http://doi.org/10.1016/j.ceb.2009.01.003

53. Pontes, O. *et al.* Natural variation in nucleolar dominance reveals the relationship between nucleolus organizer chromatin topology and rRNA gene transcription in *Arabidopsis*. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 11418–11423 (2003). http://doi.org/10.1073/pnas.1932522100

54. Gromicho, M. & Collares-Pereira, M. J. Polymorphism of major ribosomal gene chromosomal sites (NOR-phenotypes) in the hybridogenetic fish Squalius alburnoides complex (Cyprinidae) assessed through crossing experiments. *Genetica* **122**, 291-302 (2004).

55. Birchler, J. A. & Veitia, R. A. The Gene Balance Hypothesis: From Classical Genetics to Modern Genomics. *The Plant Cell* **19**, 395–402 (2007). http://doi.org/10.1105/tpc.106.049338

56. Birchler, J. A. & Veitia, R. A. Gene balance hypothesis: Connecting issues of dosage sensitivity across biological disciplines. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 14746–14753 (2012). http://doi.org/10.1073/pnas.1207726109

57. Collares-Pereira, M. J., Matos, I., Morgado-Santos, M. & Coelho M, M. Natural Pathways towards Polyploidy in Animals: The Squalius alburnoides Fish Complex as a Model System to Study Genome Size and Genome Reorganization in Polyploids. *Cytogenet Genome Research*. **140**, 97-116 (2013).

58. Machado, M. P., Matos, I., Grosso, A. R., Schartl, M. & Coelho, M. M. Non-canonical expression patterns and evolutionary rates of sex-biased genes in a seasonal fish. *Molecular Reproduction & Development* **83**, 1102-1115 (2016). https://doi.org/10.1002/mrd.22752

59. Pertea, G. *et al*. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics* **19**, 651–652 (2003). PMID: 12651724

60. Conesa, A. *et al.* Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674-3676 (2005). https://doi.org/10.1093/bioinformatics/bti610

61. Li, R. *et al*. SNP detection for massively parallel whole-genome resequencing. *Genome Research* **19**, 1124–1132 (2009). http://doi.org/10.1101/gr.088013.108

Acknowledgements

We thank Susanne Kneitz and to Ana Rita Grosso for helping with the bioinformatics data treatments and discussions. This work was funded by Project PTDC/BIA-BIC/110277/2009 to MMC and by a PhD grant (SFRH/BD/61217/2009) to IM, both from the Portuguese National Science Foundation, Fundação para a Ciência e a Tecnologia.

Authors' contributions

Conception and design by IM, MMC and MS. Fish sampling, experimental crosses and samples possessing by IM and MPM. Analysis data by IM and MPM. Data interpretation by IM. Drafting the article by IM. Revising the article by MMC, MS and MPM. Final version approved by all authors.

Competing interests

The authors declare no competing interests.

Supplementary Information and Data availability:

All additional files and datasets supporting this article are available through the figshare repository (doi: xxxxxx; https://figshare.com/s/c03974866dbd92b5a24d).

Datasets: *de novo* assembled transcriptome sequences; functional annotation of transcriptome sequences; SNP calling for all libraries.

Appendices: Table S1 and Table S2.

Supporting Information: Artificial grouping of transcripts organized according to genome-specific silencing (from i to viii).

Figures:

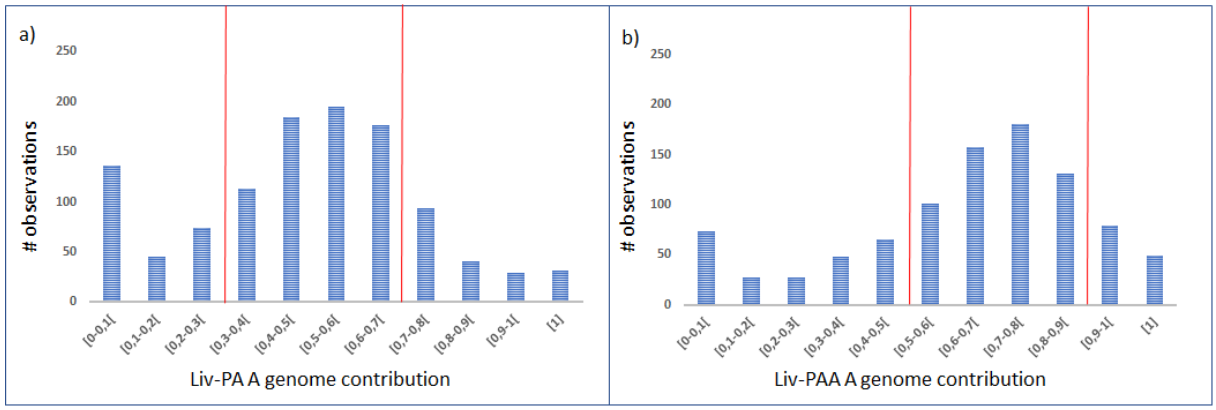


Figure 1: Distribution of transcripts according to A genome contribution to the overall gene specific transcription in liver samples. Distribution in a) liv-PA library and b) liv-PAA library. Vertical red lines represent the considered boundaries for balanced allelic expression.

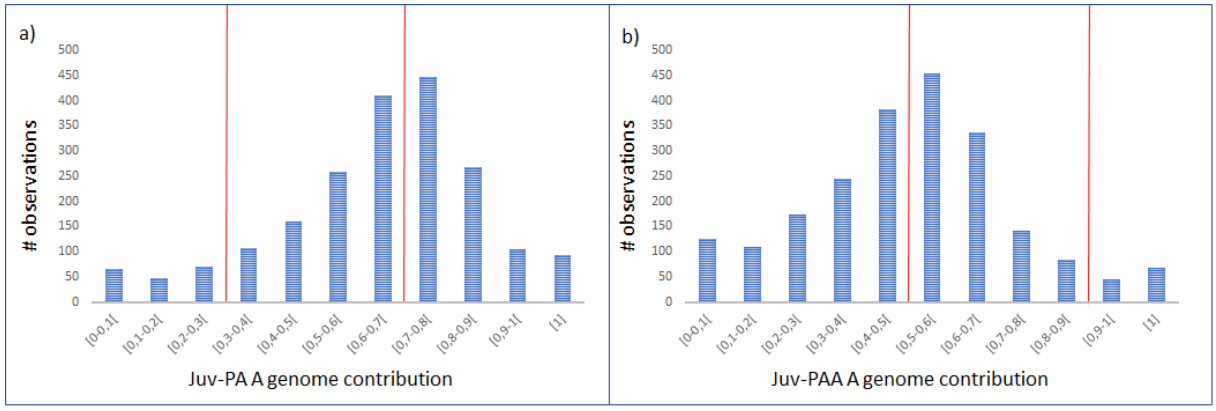


Figure 2: Distribution of transcripts according to A genome contribution to the overall gene specific transcription in juvenile samples. Distribution in a) juv-PA library and b) juv-PAA library. Vertical red lines represent the boundaries considered for balanced allelic expression.

Tables:

Table 1: Genome specific expression fractions in livers of diploid and triploid *S. alburnoides* hybrids.

	BHE	HEB	HEB(P)	HEB(A)	MGE(P)	MGE(A)	Total
Liv-PA	764 (68%)	357 (32%)	255 (23%)	102 (9%)	136 (12%)	32 (3%)	1121
Liv-PAA	649 (69%)	289 (31%)	240 (26%)	49 (5%)	73 (8%)	49 (5%)	938

BHE- Balanced homoeolog expression; HEB-Homoeolog expression bias; HEB(P)-Homoeolog expression bias towards P genome; HEB(A)-Homoeolog expression bias towards A genomic complement; MG(P)-Monogenomic expression of P alleles; MG(A)-Monogenomic expression of A alleles.

Table 2: Genome specific expression fractions in juveniles of diploid and triploid *S. alburnoides* hybrids.

	BHE	HEB	HEB(P)	HEB(A)	MGE(P)	MGE(A)	Total
Juv-PA	1386 (68%)	653 (32%)	184 (9%)	469 (23%)	66 (3%)	94 (5%)	2039
Juv-PAA	1064 (49%)	1105 (51%)	1038 (48%)	67 (3%)	125 (6%)	67 (3%)	2169

BHE- Balanced homoeolog expression; HEB-Homoeolog expression bias; HEB(P)-Homoeolog expression bias towards P genome; HEB(A)-Homoeolog expression bias towards A genomic complement; MGE(P)-Monogenomic expression of P alleles; MGE(A)-Monogenomic expression of A alleles.

Table 3: Functional enrichment in gene ontology (GO) terms and KEGG pathways of A and P monogenomic expressing (MGE) gene groups in liver and juveniles' libraries irrespective of ploidy level.

	MGE	Cat.	Term	#	Benj.
Liver	**P**	CC	Intratracellular part	28	1,10E-02
		CC	cytoplasm	21	1,90E-02
		CC	intracellular	28	3,10E-02
	A		NS		
Juvenil	**P**		NS		

	BP	translation	10	8,60E-04
	BP	peptide biosynthetic process	10	4,80E-04
	BP	organonitrogen compound biosynthetic process	12	5,90E-04
	BP	amide biosynthetic process	10	5,00E-04
	BP	peptide metabolic process	10	6,50E-04
	BP	organonitrogen compound metabolic process	13	1,30E-03
	BP	cellular amide metabolic process	10	1,60E-03
	BP	cellular protein metabolic process	17	1,50E-02
	BP	protein metabolic process	18	2,30E-02
	CC	cytosolic part	8	6,40E-06
A	CC	cytosolic ribosome	6	8,30E-05
	CC	ribosome	7	9,40E-05
	CC	cytosolic large ribosomal subunit	5	1,30E-04
	CC	ribonucleoprotein complex	9	2,10E-04
	CC	intracellular ribonucleoprotein complex	9	2,10E-04
	CC	ribosomal subunit	6	1,90E-04
	CC	large ribosomal subunit	5	3,50E-04
	CC	intracellular part	30	3,80E-04
	CC	cytoplasm	23	7,80E-04
	CC	cytoplasmic part	18	8,00E-04

CC	cytosol	8	2,50E-03
CC	intracellular	30	2,70E-03
CC	intracellular non-membrane-bounded organelle	12	4,50E-03
CC	non-membrane-bounded organelle	12	4,50E-03
CC	macromolecular complex	16	7,10E-03
CC	intracellular organelle	23	2,80E-02
CC	organelle	23	3,20E-02
MF	structural constituent of ribosome	8	3,70E-05
MF	structural molecule activity	9	7,20E-04
MF	rRNA binding	4	4,00E-03

(BP) Biological process, (MF) molecular function, (CC) cellular component, (NS) no significantly enrichment, (#) number of transcripts, (p) Benjamini corrected p-value. GO enrichment analysis was performed considering all levels of classification of terms.

Table 4: Functional enrichment in gene ontology (GO) terms and KEGG pathways of A and P monogenomic expressing (MGE) gene groups in diploid (PA) and triploid (PAA) libraries irrespective of the source tissue type.

	MGE		Term	#	p
PA	P	BP	ATP biosynthetic process	4	8,00E-02
		CC	respiratory chain	4	3,90E-02
	A	BP	peptide biosynthetic process	3	4,50E-02
		BP	peptide metabolic process	3	3,80E-02
		BP	cellular amide metabolic process	3	4,10E-02
		MF	structural molecule activity	4	1,90E-03
		MF	structural constituent of ribosome	3	7,40E-03
PAA	P	BP	carbohydrate derivative biosynthetic process	8	4,50E-02
	A	BP	translation	5	1,00E-02
		BP	peptide biosynthetic process	5	5,50E-03
		BP	amide biosynthetic process	5	5,20E-03
		BP	peptide metabolic process	5	4,90E-03
		BP	cellular amide metabolic process	5	7,20E-03
		BP	organonitrogen compound biosynthetic process	5	1,60E-02
		MF	structural constituent of ribosome	4	2,60E-03
		MF	structural molecule activity	4	1,80E-02
		KE	Ribosome	3	2,80E-02

(BP) Biological process, (MF) molecular function, (CC) cellular component, (KE) KEGG pathway, (NS) no significant enrichment, (#) number of transcripts, (p) Benjamini corrected p-value. GO enrichment analysis was performed considering all levels of classification of terms.

Table 5: Transcripts with consistent P and consistent A monogenomic transcriptional contribution regardless of sample type and ploidy level.

MGE	Unigene ID	Ref. Gene ID	Ref. Sequence	Symbol	Definition
P	Unigene101062	gi|318054652	NP_001187754.1	NDUFC2	NADH dehydrogenase (ubiquinone) 1 subunit c2 [Ictalurus punctatus]
	Unigene114185	gi|41053742	NP_957180.1	GCDH	Glutaryl-CoA dehydrogenase a [Danio rerio].
	Unigene2212	NaN			
	Unigene43419	gi|47087309	NP_998647.1	ABCD3	ATP-binding cassette sub-family D member 3 [Danio rerio]
	Unigene45629	gi|41055873	NP_957287.1		Uncharacterized protein LOC393968 [Danio rerio]
	Unigene48654l	gi|47550715	NP_999871.1	HNRNPA0	Heterogeneous nuclear ribonucleoprotein A0b [Danio rerio]
A	Unigene100595	gi|18859307	NP_571384.1	Ran	GTP-binding nuclear protein Ran [Danio rerio]
	Unigene101028	gi|18858719	NP_571660.1	FTH1	Ferritin heavy chain [Danio rerio]
	Unigene114523	gi|47523975	NP_998887.1	SLC25A3	Solute carrier family 25 member 3 [Danio rerio]
	Unigene122169	gi|225715740	gb|ACO13716.1	TIMM8A	Mitochondrial import inner membrane translocase subunit Tim8 A [Esox lucius]
	Unigene134826	gi|51010975	NP_001003447.1	RPL15	60S ribosomal protein L15 [Danio rerio]
	Unigene140252	gi|124300811	dbj|BAF45901.1	RPS13	Ribosomal protein S13 [Solea senegalensis]
	Unigene146647	gi|55250139	gb|AAH85596.1		Zgc:153867 protein [Danio rerio]
	Unigene23269	gi|47086529	NP_997925.1	RPL17	60S ribosomal protein L17 [Danio rerio]

MGE- monogenomic expression; UniGene ID from the *de novo* transcriptome assembly of *S. alburnoides* complex; Ref. Gene ID- Gene ID

CHAPTER 6

Supplementary data

Table S1. Previously constructed and sequenced libraries that have been used in this study.

Libraries	Sequencing information	Repository	Accession
liv-AA	*Matos et al., 2015*	ArrayExpress	E-MTAB-3174
liv-PP			
liv-PA			
liv-PAA			
juv-AA	Matos et al., 2015	ArrayExpress	E-MTAB-3174
juv-PA			
juv-PAA			
gonF-PP	*Machado et al., 2016*	ENA	PRJEB9465
gonM-PP			
brainF-PP			
brainM-PP			
brain-AA	*This study*	ENA	PRJEB278332
gon-AA			

Table S2. Statistics of assembly quality for *S. alburnoides* complex transcriptome.

	Total Lenght (bp)	# Contigs	Mean Lenght (bp)	Longest Contig (bp)	Shorter Contig (bp)	N50 (bp)
S. alburnoides complex transcriptome	333,800,115	263,357	1267	14300	150	100% between 0%~5%

CHAPTER 7

GENERAL DISCUSSION

7.1. The significance and role of allelic silencing in allopolyploid fish.

Allelic silencing (AS) found by Pala *et al*., (2008) in triploid *S. alburnoides*, together with targeted expression level comparisons between diploids and triploids of this fish species, lead to the disruption of the theoretical expectation that an increase in copy number of all chromosomes would have a proportional dosage effect in the expression level of all genes. Despite that AS phenomena have been implicated in the achievement of viable gene product amounts in *S. alburnoides* allopolyploids, the extension, importance and consequences of the phenomena in animal allopolyploids needed to be further explored.

7.1.1. Mosaicism as an alternative possibility for the allelic silencing quiz.

Organs are composed of several tissues, and a tissue is composed of cells with a common structure and function. Thus, in an organ, different cell types with different physiologies and embryonic origins are arranged together. When part or the whole organ is used as source of RNA, there are several cell types contributing unevenly to the total RNA extracted. In an allopolyploid context this is even more relevant, mostly when comparing organ expression between individuals with different ploidy and genomic constitutions. As such, the detection of expression differences between individuals and/or between organs could be the result of mosaicism between organs and within an organ. In fact, ploidy mosaicism has been for many years, established and documented to occur in vertebrates (Dawley and Goddard; 1988), and it appears often associated with allopolyploidy (Lamatsch *et al*., 2002; Janko *et al*., 2007; Bickham *et al*., 2009).

Very compelling in its simplicity, the hypothesis of ploidy mosaicism in *S. alburnoides* complex, and its possible implications in the allele expression imbalance

outcome observed, have been addressed at chapter 2 before more complex, costly and time-consuming paths have been taken.

To that end, flow cytometry and cell sorting protocols were developed for this system. Ploidy status evaluation in different organs and sorting of more homogenous cellular and transcriptional samples was done. Liver and kidney cell suspensions of diploid and triploid *S. alburnoides* were analysed, and ploidy mosaicism confirmed in 10% of the cases. Nevertheless, the influence of this phenomenon on the detection of variable allelic expression profiles of ubiquitously expressed genes was excluded. For several *S. alburnoides* PAA individuals, for which ploidy status as non-mosaics had been assessed, the expression pattern of three genes *(rpl8*, *gapdh* and *β-actin*) was determined. In some PAA cell samples, where P allele presence was definitively confirmed, there was still no detection of it in the transcription product (for some genes and in some organs) as previously observed and discussed by Pala *et al*. (2008). Also, an occurrence frequency of 10%, points towards diploid-triploid mosaicism as a non-regular component of the genetic system of this complex.

Interestingly, mosaicism was not detected when the sample source was the blood (discussed in chapter 2). Due to known practical reasons, blood was the elected tissue to assess the ploidy status of the *S. alburnoides* individuals in all previous studies (Próspero and Collares-Pereira, 2000; Gromicho and Collares-Pereira, 2007). This can explain why mosaicism was not earlier reported in this complex.

7.1.2. Expanded observation of gene copy silencing to other allopolyploid fishes.

The exclusion of ploidy mosaicism as ubiquitous cause for the variable allele specific expression patterns observed for housekeeping genes between tissues of the same individual, held together the assumption that a functional and meaningful mechanism of gene-regulation involving allelic silencing was operating the in triploid *S. alburnoides*. Nevertheless, if this mechanism was exclusive to this complex was still an open question.

At chapter 3, a transcriptomic assessment of the allelic expression patterns in livers of triploid medaka fish (*Oryzias latipes*) incorporating haplomes from three different medaka strains, have been accomplished. Also, at chapter 5 of this thesis the

allelic expression patterns for three target genes, in four somatic tissues, were analyzed in different triploid *Poecilia formosa* genomic contexts. With that, it was showed that allelic silencing also occurs in other fish allopolyploid situations besides in the naturally occurring triploid *S. alburnoides*. It occurs quite frequently in two distinct laboratory produced tri-genomic hybrid (TGH) configurations of *P. formosa* individuals and in engineered TGH individuals of *Oryzias latipes*.

It is interesting to note that all the situations where AS has been so far detected in allopolyploid vertebrates are cases of allopolyploidy being establish *de novo* at the moment of each individual conception. In those organisms the cells may have to "adapt immediately" to prevent catastrophic genomic shock, so the occurrence of AS can possibly be a fast way to reach a feasible coexistence and viable regulatory interactions of distinct haplomes. Nevertheless, it was showed that an abrupt ploidy and heterozygosity increase in fish do not mandatorily involves the occurrence of AS, as it is implied by the absence of AS in TGH *S. alburnoides*.

7.1.3. Molecular mechanisms intervenient or responsible for allelic silencing.

After establishing that AS occurs in other allopolyploid fish, it was important to address the molecular mechanisms underlying its occurrence.

In *S. alburnoides*, the only previous attempt in that direction was concerning the intervention of miRNA-linked mechanisms (Inácio *et al*., 2012), but other mechanisms are equally deserving of consideration.

For several reasons, discussed at chapter 5, CpG methylation seemed a good starting point, and one of the goals of this work was to look for a link between AS occurrence and the degree of total DNA methylation, in *S. alburnoides* and *P. formosa*.

It was found and presented at chapter 5, that the global methylation levels do not correlate to the AS status of each biotype. Also, the levels of DNA methylation were non-linearly related to the ploidy level in each tested allopolyploid series and the results did not show a linear correspondence between higher levels of heterozygosity and higher or lower levels of DNA methylation. Hence, the pattern of global methylation found do not fit the linear expectation of higher methylation levels to be

found in biotypes with higher AS incidence. Also, nor it clarifies the reasons for the different AS patterns found between *S. alburnoides* and *P. formosa* allopolyploid complexes. The hypothesis of a "functional diploidization" of triploids undertaken by massive methylation was not sustained. Nevertheless, it must be considered that methylation mechanisms operating at the single-locus scale may be occurring but could be undetected at the level of this analysis due to a dilution effect in the global methylation levels.

Other mechanism than DNA methylation that possibly intervene in the occurrence of allele expression bias and AS was tackled in this thesis. In view of the nucleotidic differences found between Poecilia homoeolog putative promoter regions for 3 housekeeping genes (chapter 5), the influence of different promoter strengths in the occurrence of homoeolog expression bias (HEB) and/or AS seems reasonable. In the cells of TGH *P. formosa* individuals, at each locus, three different sequences are working simultaneously as promoter of expression of each gene. As each of the different parental derived sequences can work more or less effectively as the docking site for polymerases and transcription factors originated from the other two homoeologs, different transcriptional outputs from each allele at each locus are easily expected. So, as upon any hybridization, with or without ploidy rise, homoeologous genes bring together their accompanying regulatory machinery and promoter strengths, the occurrence of any level of HEB is expected to be a common scenario, being dependent on the particular genomic context found in each hybrid.

7.1.4. Genomic context driving the patterns of allelic silencing.

Genomic context is a factor that must be considered to explain allele expression patterns in allopoliploids in general, but that have been already pointed as relevant to the occurrence of AS in *S. alburnoides* complex (Pala *et al*, 2010), conclusion that was also sustained by the results of this thesis (chapters 2 and 5).

In *S. alburnoides* complex, consistent differential patterns of gene expression have been previously found in triploid individuals with different genomic composition (Pala *et al*., 2010). Namely, it was found that the presence of C or P genomes in *S. alburnoides* biotypes result in substantial difference in genome-specific allele usage

and AS detection. However, this phenomenon could not be generally connected to the simultaneous presence of P and A genomes, since it could be a population-specific feature. At the inception of that observation, Pala *et al*., (2010) collected the C-containing specimens from two distinct northern drainages while the P-containing individuals were all from the same river -Sorraia River of Tejo basin. Therefore, at chapter 2 the study of the allelic expression patterns of P-containing *S. alburnoides* biotypes was expanded to other rivers of Tejo basin, and to other southern basins along the range of sympatry with *S. pyrenaicus*. It was found that the absence of P allele expression in some samples is persistent in the rivers where *S. pyrenaicus* is sympatric with *S. alburnoides*, but much less conspicuous than previously reported (Pala *et al*., 2008) for the Sorraia River. Consequently, it was established that the allele specific silencing expression patterns previously detected in a narrow geographic range are not local restricted, but rather pervasively related to P and A genomic interactions in the triploid PAA genomic configuration. Those findings corroborate that in *S. alburnoides* complex, the manifestation of the genomic stress, in the figure of AS, is dependent on what genomes, and of each genome intrinsic characteristics, are brought together in the same cell nucleus. The absence of detection of AS in any of the analysed tissues of TGH individuals with PQA genomic compositions (chapter 5) further supports the previous conclusion that different genome combinations lead to different mechanisms of how to cope with genomic shock, since the replacement of one A haplome for a Q haplome leads to a different outcome concerning AS occurrence. Nothing is known regarding the effect of the interaction of Q genome with A genome in natural *S. alburnoides* configurations, neither the effect of P and Q over each other without A genome influence. However, since diploid and polyploid hybrid configurations of Q genome with A genome occur naturally and the individuals are fertile and proliferate (Sousa-Santos *et al*.,2006), Q genome apparently have as good functional "affinity" with the A genome as P and C genomes have. On the other hand, if it leads to AS occurrence in QAA individuals, as in happens in PAA biotype was not tested and it is not possible to anticipate.

Also, within the *P. formosa* allopolyploid complex, concurrent presence and absence of AS in triploid individuals with different genomic compositions have been

detected. While for two different TGH *P. formosa* biotypes (mlb and mls) AS occurrence was quite frequent, in natural allotriploid (mml) genomic configurations it was not detected.

In summary, it was corroborated that the different expression patterns and the heterogeneity of allelic silencing found to occur in allopolyploid fish complexes is related to the specific genomic combination of haplomes.

7.1.5. Dosage effects but not by allele copy silencing.

The TGH individuals produced in the scope of this thesis proved to be of great utility, mostly because they offered the opportunity to distinguish the expression contribution of three different alleles in each of the three studied fish systems. All the naturally occurring allotriploid *S. alburnoides* where allelic silencing has been detected before (Pala *et al*., 2008; Pala *et al*., 2010) were carriers of a duplicated genomic set from one parental species and an unpaired genomic set from another parental species. In the cases when both parental contributions were detected, that situation did not allow to perceive if all the 3 genome copies were contributing to overall expression or if one allele (irrespective of each one) was being silenced. As for any of the TGH situations analyzed a consistent silencing of one of the 3 alleles was observed, the occurrence of a global "functional diploidization" by allele copy silencing was not supported. The most illustrative example is the case of the TGH *Oryzias latipes*, where a high throughput semi-quantitative approach was followed, and allele specific expression (ASE) was assessed for 4282 transcripts (chapter 3). For the vast majority of those, expression from all 3 alleles was detected. Nevertheless, at each locus the overall expression of the 3 alleles was similar to the expression levels found for the diploid parental state. This case is a clear indication of gene expression regulation with a dosage compensation effect, but not attained by allele copy silencing.

7.2. Transcriptomic insights on vertebrate allopolyploid gene expression

Differential allele specific expression (ASE) is a far more comprehensive phenomenon then allelic silencing (AS), that is only an extreme manifestation of ASE regulation (Grover *et al*., 2012; Yoo *et al*., 2013). Despite useful and guiding, the AS

data discussed so far and mostly based on isolated target genes, is far from sufficient to lead to a satisfactory understanding of global mechanisms of allele dosage compensation and of the implicated gene interaction networks.

To better understand the impact of allopolyploidization on a molecular genetic level, to study allele specific expression on a genomic wide scale is mandatory.

7.2.1. Novel method for high throughput allele specify expression.

Tools to assess ASE in diploid biological systems were already available and are becoming more reliable by the year (Shen *et al*, 2012, Shen *et al*., 2013). However, for non-diploid systems, assess allele-specific gene expression on a large scale is still a technical challenging problem, with limited bioinformatics resources available. In a fruitful partnership with the Molecular Biosciences group, Chemistry and Biochemistry department of Texas State University, USA, a method for determining ASE in polyploid organisms from RNA-seq data was developed in the scope of this theses and presented at chapter 3 (Garcia *et al*., 2014).

One of the bioinformatics challenges for ASE analysis in polyploids is to determine the origin of homeologs (Mcelroy *et al*., 2017). This challenge was addressed at chapter 3 through identification of diagnostic single nucleotide polymorphisms (dSNPs) that differentiate homeolog origin amongst different parental genomes. Experimentally produced artificial triploid medaka, composed of three different haplomes were specifically produced for that purpose.

General dSNPs approaches have been successfully employed before for diploid organisms (Skelly *et al*., 2011; Tang *et al*., 2011 and Zhai *et al*., 2013), but when more than 2 parental genomes are present, the bioinformatics tools available were insufficient. So, in the scope of this thesis a software tool set was developed in a way to be applicable to organisms of any ploidy level status and it was published (chapter 3).

Despite this methodology prove to be effective to determine allele-specific expression in polyploid organisms on a large scale, it is important to stress some restrictions. Limitations to this SNP-based approach are imposed by the dependence on genomic regions harboring dSNPs between parental genomes, which requires both

a certain level of genetic divergence as well as extensive genomic resources, which are still not available for many species.

7.2.2. Transcriptomic basis for a deeper study of gene expression in *S. alburnoides*.

Based on the analysis of the behavior of individual genes in the *S. alburnoides* fish complex, the fruitful efforts of Pala *et al* (2008; 2010) to clarify the impact of hybridization and polyploidization processes on genome regulation and expression have raised several questions. Those questions came to stress the necessity for wide-scale analysis of gene behavior throughout allopolyploid genomes. To do so, reasonable amount of sequence data must be available, as for medakas and like it is starting to be for amazon molly's. But for the *S. alburnoides* fish complex, there was no high throughput genomic or transcriptomic resources. So, in the scope of this thesis the first RNA-seq data on diploid and triploid *S. alburnoides* specimens and parental forms (from whole juvenile fish and from adult livers) was generated and made available in ArrayExpress (accession number E-MTAB-3174). (Chapter 4). With clear advantages over previous existing approaches, RNA-seq allows for mapping and quantifying complex transcriptomes (Wang *et al*., 2009). In the case of *S. alburnoides* complex it allowed to determine the primary sequence of transcripts (chapter 4; Machado *et al*., 2016), identify polymorphisms between several genomotypes involved in the complex (chapter 6), and determine relative abundances of each transcript within total RNA samples and between samples (chapter 5, chapter 6 and Machado *et al*., 2016). The data generated by Matos *et al*., (2015) at chapter 4, together with other RNA-Seq data sets that have been generated, lead to the production of the first "*S. alburnoides* complex" reference transcriptome (chapter 6).

7.2.3. Allele specific quantification in the *S. alburnoides* – extension and context.

At chapter 6 it was applied in a naturally occurring allopolyploid vertebrate a high throughput approach to the study of homoeolog specific expression. A main objective was to determine the relative contribution of each genome copy present in the *S. alburnoides* hybrids to the overall expression of each gene. Also, it was important

to identify and quantify the occurrence of allelic silencing at the transcriptomic scale, accessing if it is happening globally or sporadically throughout the triploid *S. alburnoides* transcriptome, and randomly or preferentially towards any of the allele copies.

It was observed that in both diploid and triploid *S. alburnoides* forms, for the vast majority of genes, there is co-expression of alleles from both P and A genomes, and only a very small percentage of genes presented allelic silencing (AS). So, no major and/or systematic silencing of one of the intervenient genomes was identified in triploids, as it would be if AS was a prime mechanisms of dosage compensation operating in *S. alburnoides*. Nevertheless, for the genes co-expressing P and A alleles, a significant percentage was found to be strongly affected by homoeolog expression bias (HEB), and more specifically by unbalanced HEB favoring P allele expression. The high incidence of HEB has been in general related to cis-regulatory divergence (Bell *et al*., 2013). Even minor regulatory variation found between homologues in conventional non-hybrids, have been found to be sufficient to significantly impact the relative expression level of alleles (McManus *et al*., 2010; Chamberlain *et al*., 2015). Accordingly, in an inter-generic hybrid as *S. alburnoides*, HEB is more probably the direct result of expected and significant variation between regulatory and/or coding sequences of P and A homoeologs, than the result of a concerted dosage compensation mechanism based on allele specific down regulation.

The fact that HEB incidence was found to be not significantly affected by ploidy, and that AS was detected not only in triploids but also in diploid *S. alburnoides*, further supported that balanced expression is not a necessity for the triploid *S. alburnoides* viability and its successful perpetuation. Also, highlights that a mechanism of specific allele down regulation does not fit as explicative scenario for the success, in terms of abundance, of the *S. alburnoides* triploid biotypes in comparison with the diploid counterparts.

On the other hand, as discussed in chapter 6, the biological significance for the HEB patterns found in *S. alburnoides* could be linked to the *S. alburnoides* specific mito-nuclear context and to the not yet well-defined genomic architecture and chromosomal context of P and A-genome NORs.

Nevertheless, it is essential to state that the evidences presented at chapter 6 do not support, but either do not refute, the existence of a direct link between AS and dosage compensation in *S. alburnoides*.

7.2.4. Dosage compensation in *S. alburnoides* – extension and context.

The mechanism of dosage compensation by allele copy silencing found in the *S. alburnoides* complex (Pala *et al*., 2008, 2010) was the inception of the attractive hypothesis that balanced expression and functional ''diploidization'' could be a necessity or an extremely relevant factor to the success and perpetuation for lower polyploid vertebrates. However, before generalizations could be made, a wider and deeper look into the occurrence of dosage compensation in *S. alburnoides* complex had to be taken. So, based on high throughput transcriptomic resources produced to that end, a first comparative quantitative transcriptomic analysis between diploid and triploid *S. alburnoides* complex individuals was performed. At chapter 5, it was showed that, despite many genes (around half) do not present a significant differential expression between diploid and triploid hybrids, the PAA gene expression level profiles are not identical to the ones of PA (or to any of the parental diploid genomotypes) ruling out the hypotheses of a global full "functional diploidization" of triploids. Yet, despite the higher gene dosage in triploids, the gene expression level profile of triploid vs diploid is not tending according to the ploidy level increase but in the opposite direction, offering an unanticipated scenario of gene expression regulation. At chapter 5 it was discarded the hypothesis of occurrence of a genome wide regulatory mechanism that would bring all genic activity of triploids to the diploid state in a "strict functional diploidization" event. Nevertheless, a considerable fraction of all triploid transcripts do suffice a strict definition for "fully dosage compensation". Consequently, it was proposed that "diploidization" might not mean that all genes are down regulated to the diploid level, but only those that need to be "diploidized" to guarantee a correct function. On the other hand, also a small percentage of the triploid transcripts are found represented strictly proportionally to gene dosage or even higher. So, the results show some transcriptional equivalence between diploids and triploids, but not as a strictly regulated and fine-tuned phenomenon. That is probably

based on a switch-like way to regulate the mRNA concentrations, with transcription turned "on" or "off" regardless of exact concentrations within a cell, but within boundaries of similar expression.

Instead of a tightly regulated phenomenon with stiff boundaries and acting without exception throughout the whole genome of the *S. alburnoides* triploids, it was found that genome expression level regulation is a more plastic process.

7.3. References

Álvarez, P. , Arthofer, W. , Coelho, M. M., Conklin, D. , Estonba, A. , Grosso, A. R., Helyar, S. J., Langa, J. , Machado, M. P., Montes, I. , Pinho, J. , Rief, A. , Schartl, M. , Schlick-Steiner, B. C., Seeber, J. , Steiner, F. M. and Vilas, C. (2015), Genomic Resources Notes. *Molecular Ecology Resources*, 15, 1510-1512. https://doi.org/10.1111/1755-0998.12454

Bell, G.D.M., Kane, N.C., Rieseberg, L. H., and Adams, K.L. 2013). RNA-Seq Analysis of Allele-Specific Expression, Hybrid Effects, and Regulatory Divergence in Hybrids Compared with Their Parents from Natural Populations. *Genome Biology and Evolution*, 5(7), 1309–1323. http://doi.org/10.1093/gbe/evt072

Bickham, J.W., and Hanks, B.G. (2009). Diploid-triploid mosaicism and tissue ploidy diversity within Platemys platycephala from Suriname. *Cytogenetic and Genome Research*, 127, 280-286. https://doi.org/10.1159/000297716

Chamberlain, A.J., Vander Jagt, C.J., Hayes, B.J., Khansefid, M., Marett, L.C., Millen, C. A., … Goddard, M. E. (2015). Extensive variation between tissues in allele specific expression in an outbred mammal. *BMC Genomics*, 16, 993. http://doi.org/10.1186/s12864-015-2174-0

Dawley, R.M., and Goddard, K.A. (1988). Diploid-triploid mosaics among unisexual hybrids of the minnows Phoxinus eos and Phoxinus neogaeus. *Evolution*, 42(4), 649-659.

Garcia, T.I., Matos, I., Shen, Y., Pabuwal, V., Coelho, M.M., Wakamatsu, Y., … Walter, R.B. (2014). Novel Method for Analysis of Allele Specific Expression in Triploid Oryzias latipes Reveals Consistent Pattern of Allele Exclusion. *PLoS ONE*, 9(6), e100250. http://doi.org/10.1371/journal.pone.0100250

Gromicho, M., and Collares-Pereira, M.J. (2007). The evolutionary role of hybridization and polyploidy in an Iberian cyprinid fish – a cytogenetic review, in Pisano, E., Ozouf-Costaz, Cm., Foresti, F., Kapoor, B.G. *Fish Cytogenetics*, Science Publishers, Enfield, U.S.A.

Grover, C.E., Gallagher, J.P., Szadkowski, E.P., Yoo, M.J., Flagel, L.E., and Wendel, J.F. (2012). Homoeolog expression bias and expression level dominance in allopolyploids. *New Phytologist*, 196, 966-971. https://doi.org/10.1038/hdy.2012.94

Inácio, A., Pinho, J., Pereira, P.M., Comai, L., and Coelho, M.M. (2012). Global Analysis of the Small RNA Transcriptome in Different Ploidies and Genomic Combinations of a Vertebrate Complex – The Squalius alburnoides. *PLoS ONE*, 7(7), e41158. http://doi.org/10.1371/journal.pone.0041158

Janko, K., Bohlen, J., Lamatsch, D., Flajshans, M., Epplen, J.T., Ráb, P., Kotlík, P., Slechtová, V. (2007). The gynogenetic reproduction of diploid and triploid hybrid spined loaches (Cobitis: Teleostei), and their ability to establish successful clonal lineages--on the evolution of polyploidy in asexual vertebrates. *Genetica*, 131(2), 185-94. https://doi.org/10.1007/s10709-006-9130-5

Lamatsch, D.K., Schmid, M., and Schartl, M. (2002). A somatic mosaic of the gynogenetic Amazon molly. *Journal of Fish Biology*, 60, 1417-1422. https://doi.org/10.1111/j.1095-8649.2002.tb02436.x

Machado, M.P., Matos, I., Grosso, A.R., Schartl, M. and Coelho, M.M. (2016). Non-canonical expression patterns and evolutionary rates of sex-biased genes in a seasonal fish. *Molecular Reproduction and Development*, 83, 1102-1115. https://doi.org/10.1002/mrd.22752

McElroy, K.E., Denton, R.D., Sharbrough, J., Bankers, L., Neiman, M., and Gibbs, H. L. (2017). Genome Expression Balance in a Triploid Trihybrid Vertebrate. *Genome Biology and Evolution*, 9(4), 968–980. http://doi.org/10.1093/gbe/evx059

McManus, C.J., Coolon, J.D., Duff, M.O., Eipper-Mains, J., Graveley, B.R., and Wittkopp, P.J. (2010). Regulatory divergence in Drosophila revealed by mRNA-seq. *Genome Research*, 20(6), 816–825. http://doi.org/10.1101/gr.102491.109

Pala, I., Klüver, N., Thorsteinsdóttir, S., Schartl, M., and Coelho, M.M. (2008). Expression pattern of antiMüllerian hormone (amh) in the hybrid fish complex of Squalius alburnoides. *Gene*, 410: 249–258. https://doi.org/10.1016/j.gene.2007.12.018

Pala, I., Schartl, M., Brito, M., Vacas, J.M., and Coelho, M.M. (2010). Gene expression regulation and lineage evolution: the North and South tale of the hybrid polyploid Squalius alburnoides complex. *Proceedings of the Royal Society B: Biological Sciences*, 277(1699), 3519–3525. https://doi.org/10.1098/rspb.2010.1071

Próspero, M.I., and Collares-Pereira, M.J. (2000). Nuclear DNA content variation in the diploid-polyploid Leuciscus alburnoides complex (Teleostei, Cyprinidae) assessed by flow cytometry. *Folia Zoologica*, 49, 53-58.

Shen, Y., Catchen, J., Garcia, T., Amores, A., Beldroth, I., Wagner, J. R., … Walter, R. B. (2012). Identification of transcriptome SNPs between Xiphophorus lines and species for assessing allele specific gene expression within F1 interspecies hybrids. *Comparative Biochemistry and Physiology. Toxicology & Pharmacology : CBP*, 155(1), 102–108. http://doi.org/10.1016/j.cbpc.2011.03.012

Shen, Y., Garcia, T., Pabuwal, V., Boswell, M., Pasquali, A., Beldorth, I., … Walter, R. B. (2013). Alternative strategies for development of a reference transcriptome for quantification of allele specific expression in organisms having sparse genomic resources. *Comparative Biochemistry and Physiology. Part D, Genomics & Proteomics*, 8(1), 11–16. http://doi.org/10.1016/j.cbd.2012.10.006

Skelly, D.A., Johansson, M., Madeoy, J., Wakefield, J., and Akey, J.M. (2011). A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome Research*, 21(10), 1728–1737. http://doi.org/10.1101/gr.119784.110

Sousa-Santos, C., Collares-Pereira, M.J., and Almada, V.C. (2006). Evidence of extensive mitochondrial introgression with nearly complete substitution of the typical Squalius pyrenaicus-like mtDNA of the Squalius alburnoides complex (Cyprinidae) in an independent Iberian drainage. *Journal of Fish Biology*, 68, S292-S301. https://doi.org/10.1111/j.0022-1112.2006.01081.x

Tang, F., Barbacioru, C., Nordman, E., Bao, S., Lee, C., Wang, X., … Surani, M.A. (2011). Deterministic and Stochastic Allele Specific Gene Expression in Single Mouse Blastomeres. *PLoS ONE*, 6(6), e21208. http://doi.org/10.1371/journal.pone.0021208

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1), 57–63. http://doi.org/10.1038/nrg2484

Yoo, M-J., Szadkowski, E., and Wendel, J.F. (2013). Homoeolog expression bias and expression level dominance in allopolyploid cotton. *Heredity*, 110(2),171-180. https://doi.org/10.1038/hdy.2012.94

Zhai, R., Feng, Y., Zhan, X., Shen, X., Wu, W., Yu, P., … Cheng, S. (2013). Identification of Transcriptome SNPs for Assessing Allele-Specific Gene Expression in a Super-Hybrid Rice Xieyou9308. *PLoS ONE*, 8(4), e60668. http://doi.org/10.1371/journal.pone.0060668

CHAPTER 8

CONCLUDING REMARKS

The main achievement of this work was to illustrate for the first time how a successful allopolyploid animal, the emblematic allopolyploid *Squalius alburnoides*, globally transcriptionally deals with the genomic stress derived from hybridization and polyploidy.

It was particularly important to clarify, concerning the *S. alburnoides* complex that an exact functional diploidization of the triploid genome does not take place, and instead of having tightly regulated boundaries, there is quantitative expression flexibility. Nonetheless, in general, a significant down regulation of gene expression in triploids does occur. This gene expression down regulation does not seem to be dependent of allele copy silencing, despite extreme homoeolog expression bias was observed to affect a significant percentage of genes in triploid *S. alburnoides*, and also in medaka triploids.

Additionally, the hypothesis of a massive methylation over triploid hybrid genomes was not sustained, both in the *S. alburnoides* complex and *P. formosa*, and a link between methylation and allelic silencing was not found in these two allopolyploid systems.

On the other hand, it was showed that allelic silencing is not a population's specific occurrence within the *S. alburnoides* complex nor a *S. alburnoides* specific phenomena. It was showed that allelic silencing in particular, but probably homoeolog expression bias in general, is a phenomenon related to the genomic context of each individual and dependent on the parental genomes brought together by hybridization. Accordingly, it was showed that homoeolog expression bias incidence is not significantly affected by ploidy, and allelic silencing was detected both in diploids and triploids. So, balanced expression does not seem to be a necessity for the triploid *S. alburnoides* viability. Also, the higher abundance of triploid *S.*

alburnoides in comparison with diploids is not explained by a mechanism of specific allele down regulation.

Other achievement of this work was the description for the first time, of the occurrence of ploidy mosaicism in the *S. alburnoides* complex, and the exclusion of this phenomena as the origin of the allele silencing detected in some tissues of some triploid *S. alburnoides* individuals.

Moreover, the RNA-seq data generated lead to the production of the first "*S. alburnoides* complex" reference transcriptome.

An additional output of this thesis was the development of a method for determining allele specific expression in polyploid organisms from RNA-seq data and the implementation of the method in a software tool set freely available.

This work reflects the complexity of allopolyploidy at the gene expression regulation level, and the results encountered highlight that finding common global rules, mechanisms or explanations that fits all allotriploid conditions might not happen as they might not exist.

The specific objectives that were put forward in the beginning of this thesis work have been essentially achieved. The obtained results, summarized above, add original data to the current knowledge on animal allopolyploidy, and clarify some aspects that have been put forward by previous investigations. Nevertheless, knowledge on any topic and at any field is static, and the questions answered with this work gave rise to several new questions and opened new avenues to future research.