GENOMIC INSIGHTS INTO SEXUAL SELECTION AND THE EVOLUTION OF REPRODUCTIVE GENES IN TELEOST FISHES

A Dissertation

by

CLAYTON MATTHEW SMALL

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2012

Major Subject: Zoology

Genomic Insights into Sexual Selection and the Evolution of Reproductive Genes in Teleost Fishes

Copyright 2012 Clayton Matthew Small

GENOMIC INSIGHTS INTO SEXUAL SELECTION AND THE EVOLUTION OF REPRODUCTIVE GENES IN TELEOST FISHES

A Dissertation

by

CLAYTON MATTHEW SMALL

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Adam G. Jones
Committee Members,	Ginger E. Carney
	Mariana Mateos
	Gil G. Rosenthal
Head of Department,	Uel J. McMahan

August 2012

Major Subject: Zoology

ABSTRACT

Genomic Insights into Sexual Selection and the Evolution of Reproductive Genes in Teleost Fishes. (August 2012)

Clayton Matthew Small, B.S., University of Idaho

Chair of Advisory Committee: Dr. Adam G. Jones

Sexual selection has long been a working explanation for the elaboration of appreciable traits in plants and animals, but the idea that it is an equally potent agent of change at the level of individual molecules is relatively recent. Indications that genes associated with reproductive biology evolve especially rapidly planted this notion, but many details about the genomics of sex remain elusive. Numerous studies have characterized rapid sequence and expression divergence of sex-related molecules, but few if any have demonstrated convincingly that these patterns exist as a result of sexual selection. This dissertation describes several genome-scale studies related to reproduction and the sexes in teleost fishes, a group of animals underexploited in regard to this topic.

Using commercial microarrays I measured the extent of sexually dimorphic gene expression in the zebrafish, *Danio rerio*. Sex-biased patterns of gene expression in this species are similar to those described in other animals. A number of genes expressed at high levels in ovaries and testes relative to the body were identified as a product of the

study, and these data may be useful for future studies of reproductive genes in *Danio* fishes.

In a second study, the recent advent of high throughput cDNA pyrosequencing was leveraged to characterize the relationships between tissue-, sex-, and species-specific expression patterns of genes and rates of sequence evolution in swordtail fishes (*Xiphophorus*). I discovered ample evidence for expression biases of all three types, and a generally positive but idiosyncratic relationship between the magnitude of expression bias and rates of protein-coding sequence evolution.

Pyrosequencing of cDNA was also used to explore the possibility that postcopulatory sexual selection drives the rapid evolution of male pregnancy genes, a novel class of reproductive molecules unique to syngnathid fishes (seahorses and pipefishes). Genes differentially expressed in the male brooding tissues as a function of pregnancy status evolve more rapidly at the amino acid level than genes exhibiting static expression. Brooding tissue genes expressed during male pregnancy have evolved especially rapidly in polyandrous lineages, a finding that supports the hypothesized relationship between postcopulatory sexual selection and the adaptive evolution of reproductive molecules.

DEDICATION

To Bruce E. and Terri J. Small, for bestowing upon me the ultimate parent's gift: unconditional encouragement.

ACKNOWLEDGEMENTS

I am greatly and humbly indebted to my advisor and committee chair, Dr. Adam G. Jones. One simply could not ask for a better, more brilliant colleague. My dissertation research would be far inferior to its ultimate culmination without the patience, countless intellectual contributions, and encouragement supplied by Dr. Jones over these past several years. I also express my utmost thanks to the other committee members: Dr. Gil G. Rosenthal, Dr. Ginger E. Carney, and Dr. Mariana Mateos. Dr. Rosenthal has been extremely enthusiastic over my work, I owe him special thanks for supporting me to play a role in his *Xiphophorus* genomics research program, and I am grateful for the opportunity to include said work in my dissertation. Dr. Carney and Dr. Mateos contributed substantially to my training at multiple levels. Their input regarding project ideas, manuscripts, and grant proposals was invaluable to me as a student, and I was also fortunate to take several outstanding courses taught by them.

During my tenure as a graduate student I was surrounded by a number of excellent and supportive peers. I especially thank my lab mate Nick Ratterman both as a great friend and a colleague, for countless scientific engagements and idea exchanges over the years. I also thank my lab mates Kenyon Mobley, Charlyn Partridge, Sunny Scobell, Kim Paczolt, Emily Rose, and Sarah Flanagan for their priceless friendship, research assistance, and ideas. Similarly, friends and Texas A&M colleagues James B. Johnson, Krithika Kumar, Seth Coleman, Zach Culumber, Rongfeng Cui, Nicole

Smolensky, Amanda Subalusky, Michael Parkes, Jody Weir, and Nick Duprey all helped to enrich and enliven my graduate training.

Academically I would like to thank a great number of people for providing assistance and resources necessary for the completion of the studies described in this dissertation. Drs. Quincy Mo, Marina Vannucci, Lisa Ellis, and Brian Perkins helped make the zebrafish microarray study possible. Olivia Ochoa and Rongfeng Cui collected *Xiphophorus* samples in Mexico and R. Cui provided a tremendous amount of assistance with data management and analysis. The pipefish genomics project was made possible by funding from the laboratory of Dr. April Harlin-Cognato at Michigan State University, and I can't thank her enough for ensuring the survival of the study. Dr. Alan Pepper was especially kind and helpful in regard to data analysis ideas and tools. Finally, the many specimens required for the study were made available through help from Dr. Kenyon Mobley, Dr. Bob Wong, Paula Branshaw, the Dallas World Aquarium, Nick Schuetz, David Mahlmann, and Rhonda Patternson.

I thank the following funding agencies for resources associated with the projects described in this dissertation: Texas A&M University, Michigan State University, the National Science Foundation, the American Museum of Natural History Lerner Gray Fund for Marine Research, and Sigma Xi.

Finally, I wish to thank my family members for years of unwavering support and for encouraging me to pursue my lifelong dream of becoming a scientist. Catherine Engelman, Terri Small, Bruce Small, Cole Small, Jean Small, Joe Knotwell, Ramona

Knotwell, Debbi Durtschi, and Michael Eisenhauer all occupy special places in my consciousness, and I am grateful for their love.

TABLE OF CONTENTS

Page

ABSTRACT iii

DEDICATION v

ACKNOWLEDGEMENTS vi

TABLE OF CONTENTS ix

LIST OF FIGURES xi

LIST OF TABLES xii

1. INTRODUCTION 1

2. A MICROARRAY ANALYSIS OF SEX- AND GONAD-BIASED GENE EXPRESSION IN THE ZEBRAFISH: EVIDENCE FOR MASCULINIZATION OF THE TRANSCRIPTOME 13

   Introduction 13
   Materials and Methods 17
   Results 22
   Discussion 35

3. SEX-, TISSUE-, AND SPECIES-SPECIFIC EXPRESSION PATTERNS INFLUENCE RATES OF SEQUENCE DIVERGENCE BETWEEN TWO HYBRIDIZING SWORDTAIL FISHES 46

   Introduction 46
   Materials and Methods 49
   Results 55
   Discussion 69

4. SEXUAL SELECTION AND THE EVOLUTION OF MALE PREGNANCY PROTEINS AMONG PIPEFISH AND SEAHORSE LINEAGES WITH DIVERSE MATING SYSTEMS 75

Introduction 75
Materials and Methods 78
Results 85
Discussion 92

5. SUMMARY 98

REFERENCES 103

APPENDIX 124

VITA 127

LIST OF FIGURES

Page

Figure 1 The Magnitude of Male- and Female-biased Gene Expression 29

Figure 2 Overlap of Sex- and Gonad-biased Gene Expression 33

Figure 3 Sex-biased Gene Expression and d*N*/d*S* in *Xiphophorus* 60

Figure 4 Tissue-biased Gene Expression and d*N*/d*S* in *Xiphophorus* 63

Figure 5 Sequence Divergence as a Function of Expression Divergence 66

Figure 6 Mating System Variation Across Four Syngnathid Fish Species 83

Figure 7 Expression Status and the Evolution of Male Brooding Proteins 87

Figure 8 Branch-specific d*N*/d*S* Estimates Among Four Syngnathid Fish Species. 88

LIST OF TABLES

Page

Table 1 Studies Relevant to the Unique Evolutionary Dynamics of Sex-related Molecules 7

Table 2 Terms Used to Describe Gene Expression Categories in this Study......... 24

Table 3 Relative Numbers of Sex- and Gonad-biased Genes in *Danio rerio* 25

Table 4 Sex- and Gonad-biased Genes Identified by Other *Danio rerio* Studies .. 26

Table 5 Top Testis-upregulated Genes in Zebrafish 31

Table 6 Top Ovary-upregulated Genes in Zebrafish........ 32

Table 7 Real-time PCR Confirmation for Seven Differentially Expressed Genes. 34

Table 8 *Xiphophorus* 454 Sequencing and *de novo* Transcriptome Assemblies ... 57

Table 9 Sex-biased Gene Expression Trends in *Xiphophorus* 58

Table 10 Magnitude of Sex-biased Gene Expression Explains Variation in d*N*/d*S* 59

Table 11 Magnitude of Tissue-biased Gene Expression Explains Variation in d*N*/d*S*. 62

Table 12 Expression Divergence Between Species Explains Variation in d*N*/d*S* ... 65

Table 13 Summary of Syngnathid Specimens Used to Generate 454 Data 79

Table 14 Syngnathid 454 Sequencing and *de novo* Transcriptome Assemblies...... 85

Table 15 Genes with Positively-selected Sites........ 90

1. INTRODUCTION

In 1871 Charles Darwin published his ultimate thesis on why males of many animal species possess extravagant, seemingly superfluous traits. Characters such as brilliant plumage in songbirds and harrowing horns in beetles, Darwin argued, evolve as a consequence of competition for access to mates. The understanding of sexual selection as an especially potent evolutionary mechanism has been in place since Darwin's original work, as is evident on page 156, Chapter 5 of The Origin of Species: "… I think it will also be admitted that species of the same group differ from each other more widely in their secondary sex characters, than in other parts of their organization…" (Darwin 1859). The rapid divergence of secondary sex characters among closely related species, the striking phenotypic differentiation between males and females of the same species, and in some cases the ability of sexual selection to drive cladogenesis, all contribute to the prominence of this theory in modern evolutionary biology (Darwin 1871; Andersson 1994; Masta and Maddison 2002; Arnegard et al. 2010).

Missing from Darwin's sexual selection synthesis, however, was the realization that competition over reproduction continues after the instance of mating itself. Indeed, an entire realm of competitive interactions involving gametes, reproductive tracts, and post-mating behaviors ultimately influence fertilization and offspring development, all of which were overlooked by Darwin and others until a century later (Eberhard 2009). It

---

This dissertation follows the style of *Molecular Biology and Evolution*.

was initially the concept of "sperm competition," the notion that sperm from multiple males ought to compete for the opportunity to fertilize a limited number of ova (Parker 1970), that inspired a deep consideration of "postcopulatory" processes among devoted students of sexual selection. A number of biological systems in which postcopulatory sexual selection may be important have recently come to light, and a variety of traits in these systems seem to have responded in a manner consistent with general sexual selection theory. For example, many studies have suggested and in some cases demonstrated that variations in morphology of male intromittent organs across a diverse array of internally fertilizing animal species have arisen as a consequence of postcopulatory sexual selection (Eberhard 2011). A well-known example of male genitalia evolution in this context is the specialized penis of the damselfly *Calopteryx maculata*, which males use to physically remove from the female reproductive tract the sperm of prior mates (Waage 1979). Other traits involved in sperm competition include attributes of the sperm themselves. Males that produce more and/or faster sperm may increase their probability of fertilization relative to competitors, as is seen in guppies (Boschetto et al. 2011). Competition after mating may also be mediated by attributes of female biology, wherein paternity among multiple mates is biased by the female via mechanisms analogous to mate choice in the pre-copulatory sense. These phenomena are known as instances of "cryptic mate choice" (Thornhill 1983) and may be facilitated by sperm storage, female egg-sperm interaction proteins, a host of female behaviors after mating and/or fertilization, and other mechanisms as of yet unknown.

Though the evidence for and understanding of postcopulatory sexual selection is dwarfed by the historical efforts made on behalf of its precopulatory counterpart, there is reason to believe that postcopulatory sexual selection also plays a potent role in generating biological variation. Most of this is evident in the impressive rates at which genital morphology diversifies among closely related species, an echo of the similar pattern long observed for more conventional secondary sex characters. This pattern may be in part attributable to selection against hybridization via reinforcement, but some have demonstrated conclusively that postcopulatory sexual selection is driving the diversification of genital morphology (reviewed in Eberhard 2011). A poignant example is a comparative study conducted by Goran Arnqvist (Arnqvist 1998), in which 19 insect clade pairs, each consisting of a monandrous clade and a polyandrous clade, were compared with respect to within-clade male genitalia diversity. Arnqvist showed that morphological divergence of male genitalia was overwhelmingly faster among clades in which females mate multiply, relative to the respective monandrous contrast clades. Some debate still exists over which particular mechanisms and models of postcopulatory sexual selection (e.g. intra- vs. intersexual effects) most broadly explain the rapid evolution of animal genitalia (Hosken and Stockley 2004), but the understanding of postcopulatory sexual selection as a driver of reproduction-related morphology is generally accepted.

Several decades after behavioral ecologists began thinking about postcopulatory sexual selection, molecular biologists interested in the evolution of reproduction noticed some intriguing patterns within their own discipline. Foremost, genes expressed with

some specificity in reproductive tissues, genes expressed differentially between the sexes, and many genes on sex chromosomes, appeared to demonstrate an elevated rate of amino acid substitution relative to the genome-wide average. This pattern was initially noted for male *Drosophila* reproductive tract proteins using two-dimensional gel electrophoresis (Coulthart and Singh 1988; Civetta and Singh 1995), and later via the comparison of protein-coding DNA sequences (Tsaur and Wu 1997). Since then numerous studies have confirmed the phenomenon in multiple animal and some plant taxa (Swanson and Vacquier 2002b; Swanson and Vacquier 2002a), by applying approaches of various scale and increasingly powerful computational methods and tools (Yang 2006). In many of these studies, for example, the authors provided evidence for positive selection having caused reproductive protein divergence. The role of positive selection in speeding up the rate of protein evolution can be distinguished from a relaxation of constraint using d*N*/d*S*, the ratio of the number of nonsynonymous substitutions per nonsynonymous site to the number of synonymous substitutions per synonymous site (Yang and Bielawski 2000). This ratio can be estimated for a set of at least two homologous protein-coding DNA sequences using a variety of codon substitution models and maximum likelihood approaches, but the general interpretation is that a d*N*/d*S* estimate less than one suggests purifying selection (evolutionary conservation of the protein), a d*N*/d*S* estimate equal to one indicates neutral evolution (absence of selection on the protein), and a d*N*/d*S* estimate greater than one implies positive selection (adaptive diversification of the protein among lineages). In most of the examples cited above, amino acid-changing DNA substitutions were shown to occur at a

higher rate than "selectively neutral" silent substitutions quite frequently during the evolutionary divergence of reproductive proteins.

Another related pattern concerns large-scale differences in transcript abundance of genes between male and female animals. Much like the obviously dimorphic secondary sex characters Darwin described, it appeared that gene expression levels could also differ quite strikingly between the two sexes, a phenomenon reviewed by Ellegren and Parsch (Ellegren and Parsch 2007). Indeed, microarray-based transcriptional profiling of the *Drosophila melanogaster* genome revealed that over half of its genes are differentially expressed between adult males and females (Ranz et al. 2003). Such extreme transcriptional differences between individuals possessing largely the same genome (save sex-specific regions in many species) seem surprising, but these gross differences make sense when one considers pleiotropy and complex gene network architecture. A few key genetic (or environmental) differences between males and females prior to and during sexual development could quite conceivably cascade into massive transcriptomic differences at maturity. Furthermore, investigators have reported that a majority of the sex-biased genes in *Drosophila* species demonstrate male-enriched expression (Arbeitman et al. 2002; Singh and Kulathinal 2005). Within males there are also many more testis-body differences in transcript abundance than there are ovary-body differences in females (Parisi et al. 2004), and the expression levels of male-enriched genes evolve more rapidly than those of female-enriched or unbiased loci (Meiklejohn et al. 2003).

To summarize these key molecular results, reproductive proteins evolve more rapidly than the genome average, gene expression is extremely sexually dimorphic in some species, and those genes that do exhibit sexually dimorphic gene expression change quickly over evolutionary time in both sequence and transcript abundance levels. An extensive list of key studies that address these patterns, annotated with focal taxa, molecule descriptions, and primary conclusions, may be found in Table 1.

A number of the authors responsible for the above work realized a natural congruence between evolutionary patterns associated with secondary sex characters and genitalia, and those patterns associated with reproductive molecules and their expression profiles. In an extension of Darwin's argument that sexual selection drives the rapid evolution of secondary sex characters, and the behavioral ecologists' argument that postcopulatory sexual selection fuels the evolution of reproductive morphology, molecular evolutionists proposed the hypothesis that sexual selection is also responsible for the rapid evolution of reproductive genes (Civetta and Singh 1995; Wyckoff et al. 2000; Swanson et al. 2001; Swanson and Vacquier 2002b). Singh and Kulathinal (Singh and Kulathinal 2005) placed emphasis on the male-specific nature of these patterns, coining the terms "male sex drive" and "genomic masculinization" to describe these phenomena.

Alternative hypotheses for the rapid and often adaptive molecular evolutionary patterns associated with reproductive processes have also been proposed. Relaxation of constraints on protein and gene expression evolution and neofunctionalization following gene duplication, two general models for accelerated molecular evolution, are cited by

Table 1 Studies relevant to the unique evolutionary dynamics of sex-related molecules.

Focal Taxa	Study Molecules	Main Conclusions	References
2 sea urchin species (*Echinometra*)	bindin (sperm fertilization gene)	Positive selection on one protein region	(Metz and Palumbi 1996)
7 abalone species (*Haliotis*)	lysin and VERL (♂ and ♀ fertilization genes)	Positive selection (lysin) and relaxed selection (VERL)	(Swanson and Vacquier 1998)
4 great ape species	3 protamines (sperm-specific histone genes)	Positive selection; rapid divergence in human-chimp lineage	(Wyckoff et al. 2000)
Caenorhabditis elegans	Genome-wide analysis	Sex-biased expression for > 12% of the genome	(Jiang et al. 2001)
2 *Drosophila* species	176 male accessory gland genes	Positive selection on 19/176 protein-coding sequences	(Swanson et al. 2001)
2 *Drosophila* species	Genome-wide analysis	Rapid, adaptive expression divergence of male-enriched genes	(Meiklejohn et al. 2003)
15 primate species	CATSPER1 (sperm calcium channel gene)	Positive selection on insertions and deletions	(Podlaha and Zhang 2003)
15 mammal species	7 male fertilization genes	Positive selection on 6/7 proteins	(Swanson et al. 2003)
2 *Drosophila* species	Genome-wide analysis	Rapid, adaptive expression divergence of reproductive genes	(Nuzhdin et al. 2004)
4 plant species (Brassicaceae)	oleopollenin gene family (7 paralogs)	Rapid, adaptive protein divergence and duplicate loss/gain	(Schein et al. 2004)
8 *Drosophila* species	169 female reproductive tract genes	Positive selection 6/169 protein-coding sequences	(Swanson et al. 2004)
8 *Drosophila* species	237 sex-biased and sex-unbiased genes	Rapid evolution of male- and female-enriched proteins	(Zhang et al. 2004)
2 *Drosophila* species	Genome-wide analysis	Most sex-biased expression is due to male adaptation	(Connallon and Knowles 2005)
Mus and *Rattus*	Genome-wide analysis	Positive selection on late-testis-development proteins	(Good and Nachman 2005)
Anopheles gambiae	Genome-wide analysis	71% of sex-biased genes are female-enriched in expression	(Hahn and Lanzaro 2005)
Homo sapiens and *Pan troglodytes*	Genome-wide analysis	Rapid evolution of X-linked relative to autosomal proteins	(Lu and Wu 2005)
3 *Drosophila* species	19 male accessory gland genes	More rapid divergence in repleta relative to melanogaster group	(Wagstaff and Begun 2005)
5 cricket species (*Gryllus*)	30 male accessory gland genes	Focal proteins evolve more rapidly than housekeeping proteins	(Andres et al. 2006)
2 from species (*Xenopus*)	Genome-wide analysis	More rapid expression divergence of female-biased genes	(Malone et al. 2006)

Table 1 Continued

Focal Taxa	**Study Molecules**	**Main Conclusions**	**References**
15 deer mouse species (*Peromyscus*)	2 egg coat genes (ZP2, ZP3)	Positive selection on both proteins	(Turner and Hoekstra 2006)
15 bird species (Galliformes)	ZP3 egg coat gene	Positive selection on protein	(Calkins et al. 2007)
12 *Drosophila* species	Genome-wide analysis	Especially rapid divergence of male-enriched and male reproductive proteins	(Haerty et al. 2007)
16 primate species	zonadhesin sperm ligand	Rapid zonadhesin evolution in polyandrous relative to monandrous species	(Herlyn and Zischler 2007)
12 primate species	2 semen coagulum genes	Rapid, adaptive evolution of SEMG1 and SEMG2; no mating system effect	(Hurle et al. 2007)
Gallus gallus and *Taeniopygia guttata*	5020 genes	Rapid divergence of Z-liked proteins relative to autosomal proteins	(Mank et al. 2007a)
Gallus gallus and *Taeniopygia guttata*	> 4000 genes	Rapid divergence of female-enriched, brain-expressed proteins	(Mank et al. 2007b)
27 salamander species (*Plethodon*)	2 courtship pheromone genes	Presence of positive selection on pheromones varies among lineages	(Palmer et al. 2007)
7 *Drosophila* species	Genome-wide analysis	Rapid expression, sequence, and turnover evolution for male-enriched genes	(Zhang et al. 2007)
2 *Caenorhabditis* species	Genome-wide analysis	Especially rapid evolution of proteins expressed in sperm	(Artieri et al. 2008)
18 rodent species	7 male reproductive genes	Positive selection on 4/7 proteins; mating system effect for 1/7 proteins	(Ramm et al. 2008)
11 rodent species	2 protamines with promoters	Association between testis mass and protein/promoter divergence of Prm 2	(Martin-Coello et al. 2009)
18 rodent species	Seminal vesicle proteome	Rapid divergence of seminal vesicle proteome complexity	(Ramm et al. 2009)
Mus and *Rattus*	704 placenta-enriched genes	Positive selection inferred for 13% of placental proteins	(Chuong et al. 2010)
14 mammal species	25 ADAM genes	Positive selection on 12 testis ADAMs; mating system association for 2/12	(Finn and Civetta 2010)
3 primate species	>10000 genes	Testis-enriched genes evolve rapidly in chimpanzee relative to human lineage	(Wong 2010)
32 mammal species	Genome-wide analysis	No relationship between mating system and overall DNA substitution rate	(Sayres et al. 2011)
12 butterfly species (*Heliconius*)	18 seminal fluid genes	Positive selection on 2 proteins; relaxed constraint in monandrous lineages	(Walters and Harrison 2011)
4 *Drosophila* species	Genome-wide analysis	Rapid evolution of male-biased, reproductive, and X-linked proteins	(Grath and Parsch 2012)

some as possible explanations for patterns specific to reproductive genes (Swanson and Vacquier 2002b). These hypotheses, however, are probably not good general explanations but rather specifically appropriate for reproductive genes with very repetitive motifs (Swanson and Vacquier 1998) or those that belong to large multi-gene families (Swanson and Vacquier 2002b), respectively. A third hypothesis for the rapid evolution of reproductive genes is diversifying selection imposed during reinforcement of speciation (Geyer and Palumbi 2003), which seems especially plausible for systems in which pre-copulatory sexual isolating mechanisms are weak or non-existent. Coevolutionary arms races between pathogens and host reproductive systems provide the impetus for a fourth hypothesis, which proposes that reproductive tracts and gamete delivery methods are especially susceptible to infection via gamete exchange (Vacquier et al. 1997). Transmission of pathogens through sex has historically received attention as an important interface between sexual and ecological selection (Hamilton and Zuk 1989), so the relevance of this concept to reproductive molecules is significant.

In general the sexual selection hypothesis is the most commonly cited explanation for rapid diversification of reproductive genes and gene expression, although it is usually divided into components arising from different mechanisms or consequences of sexual selection, namely intrasexual selection (e.g. sperm competition), intersexual selection (e.g. cryptic female choice), and sexually antagonistic coevolution (Swanson and Vacquier 2002b). Despite the popularity of the notion that sexual selection is ultimately responsible for the molecular patterns reviewed here, three key issues keep this explanation from validity as a general and satisfactory understanding within

reproductive and evolutionary biology. First, the patterns this hypothesis was formulated to explain have been confirmed in several animal and plant taxa, but there is a rather heavy bias towards several organismal groups. Model genetic systems such as *Drosophila* and *Mus* contribute to a majority of the findings, and marine gastropods and echinoderms are well represented due to their historical use as models for the biology of fertilization and reproductive isolation (Turner and Hoekstra 2008). A second issue arises from the complicated nature of gene expression patterns, the criteria used to identify genes as reproductive, male- or female-biased, and the fact that tissue specific expression itself is related to the rate of molecular evolution for a given gene (Meisel 2011). The third and most significant problem, however, is that surprisingly few published studies have directly tested for a relationship between the strength of sexual selection and the rate of reproductive molecular evolution, and no study has demonstrated unequivocally that such a relationship exits at a genomic scale (Wong 2011). The primary goals of this dissertation are to address these three empirical shortcomings and contribute to the growing knowledge about sexual selection and the genomics of reproduction using three very different groups of teleost fishes. Before I present the three main dissertation sections, however, brief descriptions of each study and the respective methodological motivations are warranted.

In Section 2 I report the results from a microarray study of sex- and gonad-biased gene expression in adult *Danio rerio*, the zebrafish. Despite the zebrafish's status as a genetic model, explicit tests for sex-biased gene expression and large-scale efforts to identify reproductive genes had not been conducted. The primary objectives of the study

described in this dissertation were to confirm whether or not transcriptome-wide sexually dimorphic gene expression occurs in zebrafish, identify putative "reproductive genes" as those up-regulated in the gonads relative to the rest of the body, and test for a major signal of male-biased gene expression in direction and magnitude, as is predicted by the "genomic masculinization" model of Singh and Kulathinal (2005). The study is rather descriptive in nature, but it addresses these relevant questions and provides reproduction-related information about the expression of over 15000 genes in *Danio rerio*.

The study presented in Section 3 takes advantage of recently developed massively-parallel cDNA sequencing in order to interrogate divergence in both transcript abundance and DNA sequences between two hybridizing swordtail fish species of the genus *Xiphophorus*. Three different tissues (gonads, sensory organs, and the remaining body) from males and females of the two species were used to generate separate transcriptome libraries composed of hundreds of thousands of short sequencing reads. The relative abundance of reads from each library permitted a quantitative evaluation of gene expression for each tissue type, and assembly of the reads into transcripts allowed for the estimation of protein-coding sequence divergence between the two species. Armed with this information, the primary goal of the study was to identify whether sex- and tissue-biased gene expression explains adaptive divergence at the amino acid level. In addition to resolving this particular issue, the study tests the general prediction that genes divergent in expression between species are also divergent at the sequence level.

In the final main section of the dissertation I present a comparative molecular evolution study designed to directly test the sexual selection hypothesis for rapid, adaptive evolution of reproductive proteins. The research described in Section 4 is especially unique for several reasons. A focus on "male pregnancy" genes expressed in the brooding structures syngnathid fishes provides a completely novel insight into the evolution of reproductive genes. Furthermore, the study includes sequence data from species with divergent mating systems. Two independent transitions from monogamous to polygamous mating systems have occurred since the divergence among these species, providing an ideal test for a statistical relationship between lineage-specific strength of sexual selection and d*N*/d*S*. This study also takes advantage of next-generation cDNA sequencing, so hundreds of orthologous coding sequences were analyzed in a massive and powerful across-species comparison.

The genomic technologies (microarrays and next-generation DNA sequencing) used to accomplish this dissertation work have only become commonplace within the last decade. Massively parallel next-generation sequencing in particular has scarcely been leveraged in efforts to explore the biological underpinnings of rapid reproductive molecular evolution. The detailed work described in the following pages, therefore, reflects an important milestone in the progress of biological research, and just one application of the unprecedented technologies revolutionizing the post-genomic era.

2. A MICROARRAY ANALYSIS OF SEX- AND GONAD-BIASED GENE EXPRESSION IN THE ZEBRAFISH: EVIDENCE FOR MASCULINIZATION OF THE TRANSCRIPTOME*

Introduction

The evolution of phenotypic differences between males and females, which are often spectacular, has been a subject of intense scrutiny since Darwin (Darwin 1871). Several well-studied, often-integrated forms of sexual dimorphism include morphological (Darwin 1871), behavioral (Breedlove 1992), and physiological (Bardin and Catterall 1981) differences. Clearly, the evolutionary mechanisms ultimately responsible for sexual dimorphism (i.e., sexual selection (Lande 1980), sex-specific ecological selection (Lande 1980), and sexual conflict (Parker and Partridge 1998)) are of great interest. However, a complete understanding of these processes is impossible without knowledge of the proximate genetic and genomic underpinnings of sex-limited phenotypes.

Several proximate mechanisms can account for the phenotypic differences between males and females. For instance, fixed genetic differences between males and females via heteromorphic sex chromosomes (Charlesworth 1991) or a sex-determination locus provide one basis for sexual dimorphism. In this case, the two sexes possess partially distinct genomes. However, phenotypic sexual dimorphism may also be

*Reprinted with permission from A microarray analysis of sex- and gonad-biased gene expression in the zebrafish: evidence for masculinization of the transcriptome, by Clayton M. Small, Ginger E. Carney, Qianxing Mo, Marina Vannucci, and Adam G. Jones. 2009. *BMC Genomics* 10:579. Copyright [2009] by Small et al.

mediated by sex differences in gene expression when a key transcript differs in abundance between males and females. These two mechanisms are by no means mutually exclusive, as sex-specific aspects of the genome result in downstream sex differences in gene expression at sex-shared loci, especially when the original sex-unique genes are highly pleiotropic (e.g. they affect multiple developmental pathways). Sexes need not have distinct genomes for sexual dimorphism to exist, however, because species characterized by environmental sex determination nevertheless maintain a considerable degree of sex-based phenotypic differentiation with respect to primary and often secondary sexual traits (Viets et al. 1993; Ewert et al. 1994; Viets et al. 1994). In these cases of non-genetic sex determination, sex differences in gene expression are obviously important sources of sexual differentiation and dimorphism.

Some interesting gene expression patterns with regard to sex have been reported over the past several years, initially in *Drosophila melanogaster* and later in other taxa (see a recent review of sex-biased gene expression by Ellegren and Parsch (Ellegren and Parsch 2007)). One observation is that of those genes that demonstrate sex-biases in expression level, more tend to be male-enriched than female-enriched (Jiang et al. 2001; Parisi et al. 2003; Rinn et al. 2004; Malone et al. 2006; Zhang et al. 2007), but there are exceptions (Hahn and Lanzaro 2005; Zhang et al. 2007). This high level of observed sexual dimorphism in gene expression is mostly attributable to differences between testis and ovary (Parisi et al. 2003). Furthermore, male-enriched genes are more divergent in their expression levels among species than are female-enriched or sex-unbiased genes (Meiklejohn et al. 2003). These patterns, in addition to the discovery that male-enriched

genes also demonstrate faster rates of DNA sequence evolution relative to female-enriched and sex-unbiased genes (Zhang et al. 2004), have been interpreted as a general signature of stronger sexual selection on males. This "male sex drive" hypothesis, formally proposed by Singh and Kulathinal (Singh and Kulathinal 2005), is consistent with findings across several animal taxa. However, additional independent tests of this hypothesis should be carried out before it is accepted as a general pattern of evolution.

In this study we take advantage of the zebrafish as a model of vertebrate reproduction to test predictions under the male sex drive hypothesis. Environment, hormones, and genetic components likely influence sex differentiation in *Danio rerio*, but the precise roles and interactions of these factors with respect to reproductive development remain unclear (von Hofsten and Olsson 2005; Wang et al. 2007). Takahashi (Takahashi 1977) originally described zebrafish gonad differentiation as a transition from a two-weeks-post-fertilization ovary-like precursor to either the mature ovary or the highly differentiated testis. This transition from ovary-like precursor to testis in males is mediated by oocyte apoptosis, which is generally complete by 29 days post-hatching (Uchida et al. 2002). More recently it has been shown that some male zebrafish exhibit few ovary-like features and lack ovary-typical gene expression during gonadal development (Hsiao and Tsai 2003). In fact, males vary dramatically in the developmental timing and abundance of ovarian features (genetic and morphological) leading up to testis formation, and there is even substantial variation within sibling groups (Wang et al. 2007). Sexual maturity in zebrafish is attained well after gonad differentiation, and usually is complete when individuals reach 23-25 mm standard

length (approximately 75 days post-hatching for domesticated strains) (Spence et al. 2008).

One advantage to zebrafish is that Affymetrix GeneChip® technology is readily available, permitting the assessment of large-scale patterns of expression in adults and their gonads. The Zebrafish Genome Array design is based on sequence information from RefSeq (July 2003), GenBank (release 136.0, June 2003), dbEST (July 2003), and UniGene (Build 54, June 2003). With approximately 14,900 transcripts represented on the array, this technology can provide a representative sample of sex differences in gene expression patterns. Our goal was to compare gene expression patterns between testes and ovaries as well as between male and female somatic tissue. A collateral benefit to these comparisons was that we were also able to identify genes within each sex that were up- or downregulated in the gonads. Under the male sex drive hypothesis, we expected more genes upregulated in males relative to females. We predicted many of these genes to be gonad specific, but also expected to find some genes expressed at different levels in the somatic tissues of males compared to females.

While our study is the first to explicitly address the male sex drive hypothesis in *Danio rerio*, several recently published microarray studies of gene expression in zebrafish have addressed aspects of sexually dimorphic gene expression and gonad specific expression patterns. In general these studies have revealed that the quantities of particular transcripts often differ significantly in adult males and females, at the level of the whole body (Wen et al. 2005), the gonads (Santos et al. 2007; Sreenivasan et al. 2008), the brain (Santos et al. 2008; Sreenivasan et al. 2008), the liver (Robison et al.

2008), and other tissues (Sreenivasan et al. 2008). However, these studies do not necessarily agree with ours on all points related to patterns relevant to the evolution of sex-biased gene expression in zebrafish, so we will return to this topic in the discussion.

Materials and Methods

Affymetrix GeneChip® preparation

We allowed eight mating pairs of wild-type (AB laboratory strain) *Danio rerio* to spawn under controlled laboratory conditions and subsequently separated the sexes for a period of five days to prevent re-mating and standardize reproductive cycles. To minimize inter-individual differences among the fish, all subjects were full siblings, between 4 and 12 months old. After sacrificing each individual by ice bath euthanasia, we quickly excised all testicular tissue from males and all ovarian tissue from females. All methods were approved by Texas A&M University's Institutional Animal Care and Use Committee (AUP2005-76). Tissues were flash-frozen in TRIzol® Reagent (Invitrogen), and total RNA isolation was performed in accordance with the manufacturer's guidelines. Following quantification and quality assessment, total RNA samples from three testis pairs, three male bodies, three ovary pairs, and three female bodies were sent to the University of Kentucky Microarray Core Facility for cRNA labeling and hybridization to 12 GeneChips® using standard Affymetrix protocols (described in the GeneChip® Expression Analysis Technical Manual). Briefly, total RNA was reversed transcribed, followed by production of biotinylated cRNA. After a fragmentation step the biotinylated cRNA was hybridized to the arrays for a period of 16 hours. The

samples were then stained with streptavidin phycoerythrin and amplified using a biotinylated anti-streptavidin antibody prior to scanning.

Absolute expression analyses

The GeneChip® Zebrafish Genome Array contains ~15,500 probe sets, each set consisting of 16 adjacent but non-overlapping probe pairs. These probe pairs are 25 bases long, each pair containing one probe (*PM*) that perfectly matches the target transcript and another probe (*MM*) that mismatches the target sequence at a single base pair. The presence of a mismatch probe is intended to control for background noise caused by hybridization of non-target molecules. To convert array image information into transcript abundance values, we employed four different "absolute expression analysis" algorithms. Each of these analysis methods was used to generate a distinct dataset from a given chip image file. We applied standard normalization procedures to raw data prior to analysis, as suggested by each respective program manual. Normalized expression values for all absolute analyses across all experimental replicates, along with other pertinent microarray details, have been deposited into the NCBI Gene Expression Omnibus (GEO) under accession number GSE14979.

GCOS

The algorithm implemented in the GCOS software package (Affymetrix), uses the one-step Tukey's biweight mean of

$$PM_i - CT_i$$

across *i* probe pairs, where *PM* is the intensity of the perfect match probe cell, and *CT* is the "contrast value" (Hubbell et al. 2002; Liu et al. 2002). *CT* is most often equal to *MM* (the intensity value of the mismatch probe cell), but if many probe pairs within a set demonstrate *MM* values larger than their corresponding *PM* values, an adjusted value is used for *CT* to eliminate the computation of negative expression values (Rajagopalan 2003). This algorithm is therefore a simple calculation based on subtracting background noise from the putative "true signal."

GC-RMA

We also used the GC-RMA (GC Robust Multi-Array Analysis) algorithm, as implemented in the microarray analysis software package GeneSpring GX 7.3.1 (Agilent). The GC-RMA algorithm is based on a linear additive model, and thus considers all arrays in a given dataset when estimating expression values for each chip, unlike the GCOS algorithm. The basic linear model is described by Wu et al. (Wu et al. 2004), and assumes that

$$Y_{gij} = O_{gij} + N_{gij} + S_{gij} \,,$$

where Y_{gij} is the *PM* intensity value for probe *j* in probe set *g* on array *i*. O_{gij} is the corresponding "optical noise" due to laser scanning errors, N_{gij} is the corresponding "non-specific binding noise," and S_{gij} is a quantity proportional to the actual abundance of target transcript in a sample (which allows for estimation of the "true" expression value). The GC-RMA algorithm uses many parameters from the observed data in all

arrays to estimate components of N_{gij} and S_{gij}, then it fits the model to calculate expression values (Wu et al. 2004).

PM-MM, PM-Only

Two additional model-based approaches, available in the analysis package dChip (Li and Hung Wong 2001; Li and Wong 2001), were also used to generate expression values. The PM-MM model assumes that for every probe set in a group of *i* arrays,

$$PM_{ij} - MM_{ij} = q_i f_j + e_{ij} \,,$$

where PM_{ij} and MM_{ij} are the perfect match and mismatch intensities for probe pair *j* in array *i*, q_i is the expression index for the probe set in array *i* (the value of interest), f_j is a coefficient that represents the relationship between probe pair *j* cell intensities and actual target concentration, and e_{ij} is the model's error term (Li and Hung Wong 2001; Li and Wong 2001; Rajagopalan 2003). Similar to GC-RMA, the PM-MM algorithm uses information from all chips in a dataset, and then the model is fit to estimate the expression value for each probe set on each chip. The PM-Only algorithm is similar to PM-MM, but the mismatch intensities are completely ignored in the model:

$$PM_{ij} = q_i f_j + e_{ij}.$$

This alternative model was created to avoid the occasional calculation of negative expression values when *MM* probe intensities are high compared to *PM* intensities (Li and Hung Wong 2001; Li and Wong 2001).

Comparative expression analyses

To compare absolute expression values between different treatment groups, detect differential transcript levels, and estimate fold changes, we conducted standard t-tests using the Cyber-T web interface (Baldi and Long 2001). This approach yielded 4 sets (one per absolute expression algorithm) of results for each of the following comparisons: male body vs. female body, testis vs. ovary, testis vs. male body, and ovary vs. female body. To control for the statistical problem of performing ~15,000 t-tests per comparison, we set a false discovery rate (FDR) of 0.05, as described by Benjamini and Hochberg (Benjamini and Hochberg 1995), for each analysis. To decide whether a gene for a given comparison was to be considered “differentially expressed,” we adopted a “strict consensus” criterion wherein the gene was required to demonstrate a significant FDR-adjusted p-value across all 4 absolute analysis datasets. This procedure is conservative, but justifiable in the name of controlling for false positives.

Real-time PCR

We used the remaining 5 male and 5 female zebrafish samples to conduct independent tests of expression bias for seven genes identified as differentially expressed by our microarray analyses. Within each of the testis-upregulated, male-enriched, and female-enriched categories we randomly chose two of the top ten most upregulated genes. We were able amplify a gene-specific PCR product for only one of the chosen male-enriched transcripts (probe set 15637.1.S1_at). Within the ovary-upregulated category, we randomly chose two of the top 200 most upregulated genes, in order to assess the

accuracy of microarray results for genes demonstrating less striking differences in expression. For each sample the same quantity of total RNA (1 μg) was reverse transcribed into cDNA using the Superscript® First Strand Synthesis Kit (Invitrogen).

We performed real-time PCR using the SYBR® Green PCR Mastermix (Invitrogen) and 2 μl of cDNA template. Reactions were run on an ABI 7700 real-time PCR apparatus (Applied Biosystems) using default analysis settings. Each individual reaction was performed in triplicate, and no-template controls were included for each primer pair to confirm amplification specificity. A dilution series including five different template concentrations was employed to facilitate the Relative Standard Curve Method (Applied Biosystems) for estimating relative mRNA levels. Primer sequences for target genes were designed using Primer Express® 3.0 (Applied Biosystems) and are available upon request. Two sets of control primers (suggested in Tang et al. (Tang et al. 2007)) were used to normalize the abundance of cDNA in each reaction. *EF1α* was used in the gonad-body comparisons, and *Rpl13α* was used in the male-female comparisons. For each comparison we calculated a 95% confidence interval about mean fold change, based on the expression level estimates across the five experimental replicates.

Results

Sex-biased gene expression

To assess the extent of sex-biased gene expression in *Danio rerio* we compared male body to female body transcript levels, and we performed a separate testis-ovary comparison. This effectively allowed us to isolate the proportion of sex-biased gene

expression attributable to differences between male and female gonads. To avoid any confusion about references to the different gene expression categories, Table 2 outlines the relevant terminology, to which we hereafter adhere. Based on our expression bias criteria, 5899 out of 15502 probe sets (38%) represented on the Affymetrix zebrafish GeneChip® demonstrated statistical testis-ovary differences, across all four absolute expression analyses, in transcript abundance. 1737 probe sets yielded an insufficient signal in all ovary and testis replicates. Of the 5899 sex-biased genes, 3387 were positively biased in males ("male-enriched"), and 2512 were positively biased in females ("female-enriched") (Table 3), consistent with the overall direction of sex-biased gene expression documented in other taxa (Parisi et al. 2003; Rinn et al. 2004; Malone et al. 2006; Zhang et al. 2007). Also represented in Table 3 are the numbers of sex-biased genes corresponding to increasingly stringent fold change criteria. From this information it is clear that the direction of sex-biased gene expression remains robust, even when genes demonstrating small sex differences in expression are not considered. Additional Files 1 and 2 (Appendix) contain lists of all male- and female-enriched genes, respectively. Other zebrafish studies have detected male- and female-enriched genes via comparison of testis and ovary (Santos et al. 2007; Sreenivasan et al. 2008).

We selected five male-enriched and five female-enriched genes from Santos et al. (Santos et al. 2007) and from Sreenivasan et al. (Sreenivasan et al. 2008) to confirm that these 20 genes fall into the same expression categories in our study (see "male-enriched" and "female-enriched" sections of Table 4). We selected these genes because they ranked at the top of their respective lists in regard to the magnitude of expression

bias. As Table 4 indicates, 18 out of these 20 major sex-biased genes from (Santos et al. 2007) and (Sreenivasan et al. 2008) are also among our list of sex-biased genes.

It is important to note that we detected no gene expression biases between male and female body tissue under our strict criteria for significance (1574 probe sets demonstrated an insufficient signal in all male body and female body replicates). If we relax our criteria by allowing statistical significance in any one of the four analysis

Table 2 Terms Used to Describe Gene Expression Categories in this Study.

Term	Explanation
Male-enriched	Genes demonstrating greater transcript abundance in the testes relative to the ovaries.
Female-enriched	Genes demonstrating greater transcript abundance in the ovaries relative to the testes.
Testis-upregulated	Genes demonstrating greater transcript abundance in the testes relative to the male body (from which the testes have been removed).
Ovary-upregulated	Genes demonstrating greater transcript abundance in the ovaries relative to the female body (from which the ovaries have been removed).

algorithms (as opposed to all four) to constitute evidence of differential expression, then we find 112 genes that are differentially expressed between male and female body tissue. This list of putative sexually dimorphic genes is included as supplementary information (Additional File 3, Appendix), but these genes are not considered in further analyses within this study. Indeed, other microarray studies of zebrafish have demonstrated sex differences in isolated organs such as the liver (Robison et al. 2008) and the brain

Table 3 Relative Numbers of Sex- and Gonad-biased Genes in *Danio rerio*. Number of sex- and gonad-biased genes (strict consensus FDR = 0.05) under increasing fold change thresholds. As the fold change criterion becomes more stringent, fewer genes are deemed differentially expressed, but the male-biased patterns remains consistent. The numbers above reflect genes that satisfy the indicated fold change thresholds across all four absolute expression analyses.

Expression bias class	No fold threshold	≥ 1.5 fold	≥ 2 fold	≥ 4 fold	≥ 6 fold
Male-enriched	3387	3219	2576	1196	728
Female-enriched	2512	2281	1684	664	413
Testis-upregulated	3002	2824	2159	925	554
Ovary-upregulated	981	842	426	0	0

(Santos et al. 2008), but according to our results, the vast majority of sex-biases in zebrafish gene expression are due to transcriptomic differences between testis and ovary. This observation is consistent with studies of other taxa in which tissue-specific contributions to sex-biased gene expression have been parsed out (Parisi et al. 2003; Parisi et al. 2004; Rinn et al. 2004).

To further examine whether the overall magnitude of sex-biased gene expression in zebrafish is greater for male-enriched genes, we compared fold change values of male-enriched genes to those of female-enriched genes. For each gene, the mean fold change estimate across all four absolute expression analysis estimates (GCOS, GC-RMA, PMMM, and PM-only) was used to represent the magnitude of expression bias. The male-enriched and female-enriched distributions of this variable are significantly

Table 4 Sex- and Gonad-biased Genes Identified by Other *Danio rerio* Studies. List of sex- and gonad-biased genes identified by other recent zebrafish studies (Li et al. 2004; Santos et al. 2007; Sreenivasan et al. 2008). The sex-biased genes are based on testis-ovary comparisons, as in our study. These genes were chosen from the above studies based on reportedly high expression bias. We screened our lists of differentially expressed genes to assess agreement with the other studies. The "fold rank" is the position each gene occupies in our lists, based on the mean of rank across the four absolute expression comparisons. *sept4*, for example, is the gene demonstrating the second-highest male-enriched expression (out of 3387 total male-enriched genes). No rank is listed if the gene failed to pass our "strict consensus" statistical criteria (see Methods). Also listed are fold change estimates from each absolute expression analysis.

Gene Name, EST accession number (if applicable)	Reference	Fold Rank (This Study)	GCOS Fold	GC-RMA Fold	PMMM Fold	PM Only Fold
Male-enriched Genes						
anti-Mullerian hormone (amh)	Santos et al. 2007	18	328.39	154.95	78.09	80.60
cyclin G2 (ccng2)	Santos et al. 2007	690	328.39	11.78	7.74	8.85
heat shock cognate 70-kd protein (hsp70)	Santos et al. 2007	-	3.64	2.63	2.20	2.63
similar to *septin 4 (sept4)*	Santos et al. 2007	2	608.87	673.38	48.17	364.57
tubulin, alpha 7 like (tuba7l)	Santos et al. 2007	19	235.72	985.76	46.06	66.53
similar to *tektin 1*, CO352798	Sreenivasan et al. 2008	3	484.87	681.64	49.65	196.92
dynein, axonemal, intermediate polypeptide 1 (dnai1), CO355627	Sreenivasan et al. 2008	45	144.47	186.74	27.69	50.98
similar to *human AKAP-associated sperm protein*, CO353327	Sreenivasan et al. 2008	83	58.05	181.15	22.55	44.38
piwi-like 1 (Drosophila) (piwil1), CO354057	Sreenivasan et al. 2008	1261	6.81	8.28	5.12	5.01
similar to *testis-specific-A-kinase-anchoring protein*, CO354405	Sreenivasan et al. 2008	4	409.79	416.74	58.48	174.20
Female-enriched Genes						
transmembrane phosphatase with tensin homology (tpte)	Santos et al. 2007	1084	3.34	4.50	3.30	3.48
RNA binding protein with multiple splicing 2 (rbpms2)	Santos et al. 2007	216	26.94	48.76	14.73	20.47
connexin 44.2 (cx44.2)	Santos et al. 2007	139	66.33	198.28	31.92	54.72
SRY-box containing gene 11b (sox11b)	Santos et al. 2007	187	30.21	79.53	22.06	23.13
cyclin B2 (ccnb2)	Santos et al. 2007	284	14.58	28.39	12.39	13.34
similar to *egg envelope glycoprotein*, CO350790	Sreenivasan et al. 2008	132	62.31	190.67	53.93	59.02
flap structure-specific endonuclease 1 (fen1), EV603088	Sreenivasan et al. 2008	-	1.11	1.17	1.10	1.10

Table 4 Continued

Gene Name, EST accession number (if applicable)	Reference	Fold Rank (This Study)	GCOS Fold	GC-RMA Fold	PMMM Fold	PM Only Fold
Female-enriched Genes						
hypothetical protein LOC556628, CO350423	Sreenivasan et al. 2008	30	273.38	1679.77	110.57	156.60
B-cell translocation gene 4 (btg4), CO349959	Sreenivasan et al. 2008	75	168.79	416.47	78.66	168.05
similar to *transcription factor IIIA*, CO349799	Sreenivasan et al. 2008	138	58.00	147.42	40.73	60.90
Testis-upregulated Genes						
zgc:162225, CO352964	Li et al. 2004	139	282.01	121.35	9.68	33.48
WD repeat-containing protein 69, CO355324	Li et al. 2004	82	59.34	175.15	21.17	41.71
zgc:158652, CO353149	Li et al. 2004	-	113.93	392.11	30.49	42.58
zgc:112008, CO352835	Li et al. 2004	176	28.11	111.65	13.66	24.40
similar to CG14551-PA, CO352954	Li et al. 2004	9	301.99	280.30	46.03	207.76
hypothetical protein LOC558005, CO355049	Sreenivasan et al. 2008	147	69.83	220.63	24.25	14.03
unknown transcript, CO355999	Sreenivasan et al. 2008	383	12.96	9.27	12.50	56.80
hypothetical protein LOC100003104, CO353145	Sreenivasan et al. 2008	13	165.59	1010.46	42.16	81.04
similar to *polyprotein*, CO355597	Sreenivasan et al. 2008	63	51.37	131.92	47.53	48.39
similar to *tektin 1*, CO353325	Sreenivasan et al. 2008	12	167.65	848.47	57.56	69.06
Ovary-upregulated Genes						
hypothetical protein LOC100001369, CO350972	Li et al. 2004	-	3.45	3.29	2.22	2.44
hypothetical protein LOC555929, CO351149	Li et al. 2004	-	2.66	5.13	2.19	3.21
similar to *novel rhamnose binding lectin*, CO350303	Li et al. 2004	-	1.17	1.21	1.14	1.20
unknown transcript, CO350393	Li et al. 2004	-	2.03	2.79	1.88	1.97
similar to *egg envelope glycoprotein*, CO350790	Sreenivasan et al. 2008	-	1.37	1.92	1.48	1.52
wu:fi40a06, CO349940	Sreenivasan et al. 2008	-	1.26	1.93	1.48	1.49
DEAD (Asp-Glu-Ala-Asp) box polypeptide 56, CO354027	Sreenivasan et al. 2008	810	2.38	2.47	1.60	1.58
hypothetical protein LOC447813, CO350110	Sreenivasan et al. 2008	-	2.02	1.76	1.02	1.02
clone MGC:55720, CO350755	Sreenivasan et al. 2008	909	2.13	2.37	1.35	1.35
retinol saturase like (retsatl), CO350808	Sreenivasan et al. 2008	-	1.75	2.49	1.79	1.85

different (Mann-Whitney U Test, $p < 0.001$), the male-enriched fold change values being greater in magnitude. Frequency distributions of male- and female-enriched genes are represented graphically in a mirrored histogram (Figure 1). Based on Figure 1, it is evident that the male-enriched gene distribution includes more "high fold change" observations than the female-enriched distribution.

Gonad-biased gene expression

It might be argued that transcripts more abundant in an organism's gonads relative to its body correspond to genes especially relevant to reproduction. In light of this, we thought it would be informative and useful to identify putative reproductive genes in *Danio rerio*. According to our criteria for differential expression, 3002 genes represented on the array were upregulated in the testes, and 2338 were downregulated (1297 probe sets yielded an insufficient signal in all testis and male body samples). 981 genes were upregulated in the ovaries, and 1399 were downregulated (1917 probe sets produced an insufficient signal in all ovary and female body samples). The numbers of differentially expressed genes decline as one imposes more stringent fold-change criteria (Table 3), and it appears that ovary-upregulated genes demonstrating high fold changes are scarce, relative to high-fold testis-upregulated genes. Complete lists of testis- and ovary-upregulated genes are included as Additional Files 4 and 5 (Appendix), respectively. Our results indicate that male, compared to female, zebrafish possess many more genes whose expression is elevated in gonads.

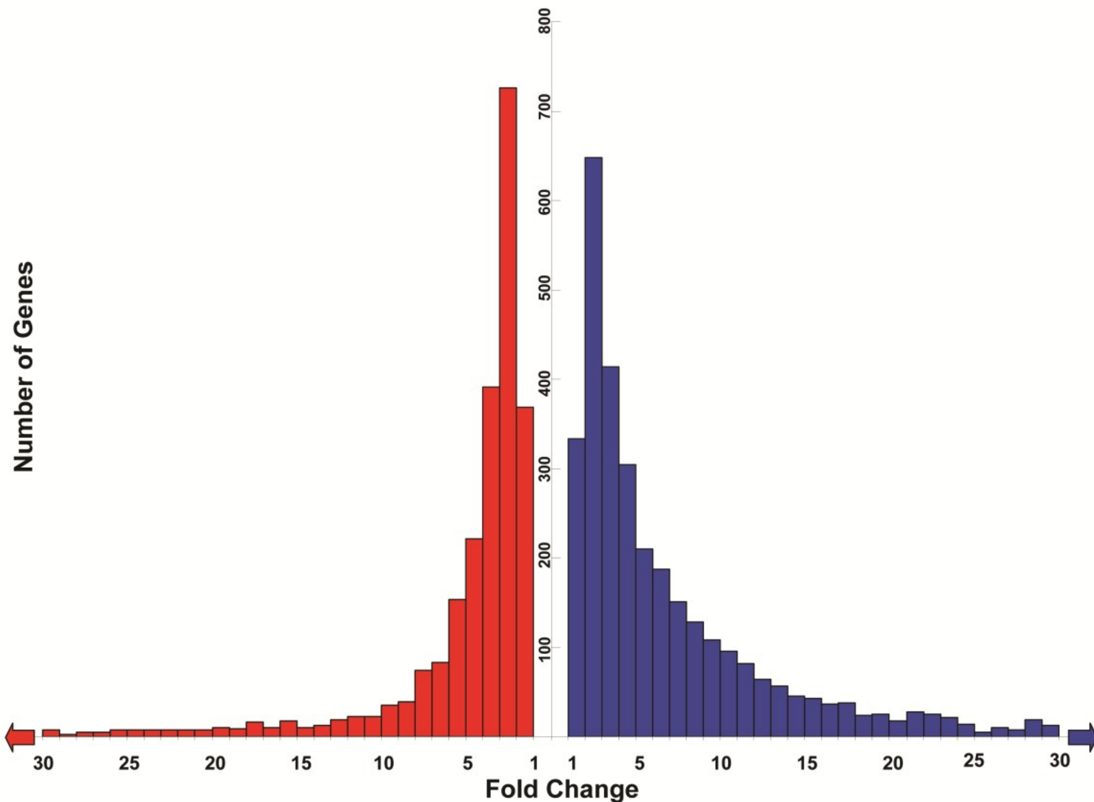

Figure 1 The Magnitude of Female- and Male-Biased Gene Expression.
Histogram showing the distributions of fold change values for female-enriched (red) and male-enriched (blue) genes. Recall that our differential expression criteria revealed 2512 female-enriched and 3387 male-enriched genes. Each observation represented in this graph is a mean across four fold change values, corresponding to the four different absolute expression analyses. Arrows at x-axis termini represent distribution tails, which are not shown. These tails (approximately 200 observations each) were omitted for ease of graphical representation, and their absence does not affect the interpretation of the histogram. Comparison of the two distributions reveals that male-enriched genes are more frequent at higher fold change intervals, relative to female-enriched genes, and a Mann-Whitney U test formally confirms higher fold change values for male-enriched genes ($p < 0.001$).

To identify which testis- and ovary-upregulated genes demonstrated the highest gonad specificity, we ranked each gene based on its average fold change ranking across each absolute expression analysis dataset. Tables 5 and 6 report the 15 highest ranking testis- and ovary-upregulated genes, respectively. For some of the genes corresponding to known or predicted *Danio rerio* mRNAs, functional annotation information is available. In some cases (Table 5) this information confirms the presumed reproductive functions of these genes. The testis-enriched gene *odf3l*, for example, codes for a structural protein (SHIPPO 1) associated with the sperm flagellum (de Carvalho et al. 2002), and may therefore be of relevance with respect to sperm competition.
For the most part, however, it is difficult to speculate on the actions of gene products that remain largely uncharacterized.

Other studies have identified genes upregulated in or specific to zebrafish gonads, based on various methods and expression criteria (Zeng and Gong 2002; Li et al. 2004; Sreenivasan et al. 2008). We selected five testis-upregulated and four ovary-upregulated genes from Li et al. (Li et al. 2004), and five testis-upregulated and six ovary-upregulated genes from Sreenivasan et al. (Sreenivasan et al. 2008) to confirm that these 20 genes fall into the same expression categories in our study (see “testis-upregulated” and “ovary-upregulated” sections of Table 4). We selected these genes because they ranked at the top of their respective lists in regard to the magnitude of expression bias. While our study agrees with these other two studies quite well in terms of testis-upregulated genes, there is rather poor agreement over ovary-upregulated genes.

Table 5 Top Testis-upregulated Genes in Zebrafish.
Fifteen highest ranking testis-upregulated genes (of 3002 total), determined by the mean of all four fold change rank values for each of the absolute expression analyses. Basic annotation is represented by a top MegaBLAST hit for each GeneChip® probe set sequence, obtained by a search of the GenBank reference mRNA database. Any supplementary functional annotation information is included if available. E-values for the above BLAST searches are all 0.0, except for *sept4* (5 e-65) and *cyp17a1* (1 e-123). Several of the probe sets listed here lack any information with respect to a described mRNA counterpart, and many correspond to hypothetical protein-coding transcripts. Three of the well-annotated transcripts (in bold text), appear to be reproduction-related.

GenBank acc. #	GenBank reference mRNA sequence	GCOS fold	GC-RMA fold	PM fold	PMMM fold
NM_001082815	similar to *septin 4* (*sept4*)	590	1162	47	161
NM_212833	zgc:56699	404	1501	40	252
BI709397	unknown. No significant BLAST hits.	254	1026	51	96
BI709397	unknown. No significant BLAST hits.	412	451	36	331
NM_199958	***outer dense fiber of sperm tail gene 3-like* (*odf3l*)**	**162**	**787**	**71**	**97**
NM_212806	*cytochrome P450, family 17, subfamily A, polypeptide 1* (*cyp17a1*)	157	368	92	184
NM_131057	***vasa homolog* (germ line development)**	**287**	**558**	**40**	**145**
NM_001100021	UPF0722 protein, *C11orf88* homolog	146	541	73	101
XM_692188	similar to CG14551 CG14551-PA	302	280	46	208
NM_001002357	zgc: 92129	349	1785	69	61
NM_001118894	***synaptonemal complex protein 1* (*sycp1*)**	**452**	**191**	**62**	**191**
NM_001007397	zgc:101797	168	848	58	69
XM_001342700	similar to predicted protein (LOC100003104)	166	1010	42	81
XM_692362	wu:fj98c04	187	401	34	185
NM_001089414	hypothetical protein zgc:162591	203	337	47	95

Table 6 Top Ovary-upregulated Genes in Zebrafish.
Fifteen highest ranking ovary-upregulated genes (of 981 total), determined by the mean of all four fold change rank values for each of the absolute expression analyses. Basic annotation is represented by a top MegaBLAST hit for each GeneChip® probe set sequence, obtained by a search of the GenBank reference mRNA database. Any supplementary functional annotation information is included if available. E-values for the above BLAST searches are all 0.0, except for *nsmce1* (2 e-152). Several of the probe sets listed here lack any information with respect to a described mRNA counterpart, and many correspond to hypothetical protein-coding transcripts.

GenBank acc. #	GenBank reference mRNA sequence	GCOS fold	GC-RMA fold	PM fold	PMMM fold
XR_044724	zgc:109744	5.2	9.8	3.4	5.5
NM_001123299	similar to CG14692-PA	5.0	8.8	3.8	4.4
XM_678859	similar to *tripartite motif protein 33*	4.7	9.6	3.4	5.5
NM_001003609	*microtubule associated serine/threonine kinase-like* (*mastl*) [associated with amino acid phosphorylation]	4.4	7.6	3.6	4.0
BM957577	unknown. No significant BLAST hits.	4.1	7.5	3.4	3.9
NM_200329	*globoside alpha-1,3-N acetylgalactosaminyltransferase 1-like 1* (*gbgt1l1*) [homologous to *mammalian ABO transferase A*]	5.0	5.6	3.6	3.8
XM_001920491	similar to *Tudor domain-containing protein 6* (Antigen NY-CO-45) (Cancer/testis antigen 41.2) (CT41.2)	4.9	7.2	2.6	4.6
NM_001017680	*F-box protein 16* (*fbxo16*)	4.9	6.2	3.1	3.3
NM_001123056	zgc:172124 [homologous to *protein kinase C, eta*]	4.4	9.0	2.6	4.3
NM_001098186	*suppressor of variegation 4-20 homolog 2* (*Drosophila*) (*suv420h2*)	3.6	12.2	3.9	5.4
NM_001020771	zgc:112481	4.3	5.3	3.1	4.0
XM_001339628	*jumonji domain containing 2A-like* (*jmjd2al*)	4.4	6.3	2.9	3.3
NM_001002551	*non-SMC element 1 homolog* (*S. cerevisiae*) (*nsmce1*)	4.1	6.1	2.6	4.3
NM_001077170	im:7162391, *nephrocystin-1*	3.9	6.5	2.8	3.3
NM_001100948	*granulito*	3.9	5.3	3.2	3.4

There is a large categorical overlap with respect to sex- and gonad-biased gene expression (Figure 2). Approximately 27% of the genes that were identified as being either male-enriched or testis-upregulated intersect. This dual categorical identity also exists for ~23% of genes that are either female-enriched or ovary-upregulated. In general, a substantial proportion of genes upregulated in the gonads of each sex are also expressed differentially between male and females.

Validation of microarray expression measurement

We used real-time PCR to confirm transcription bias in a subset of genes, representing the four different microarray expression bias categories relevant in this study (See Methods for details). Seven genes (two ovary-upregulated, two testis-upregulated, two female-enriched, and one male-enriched) were selected based on high fold change rank within each class and amenability to successful PCR amplification. A summary of the

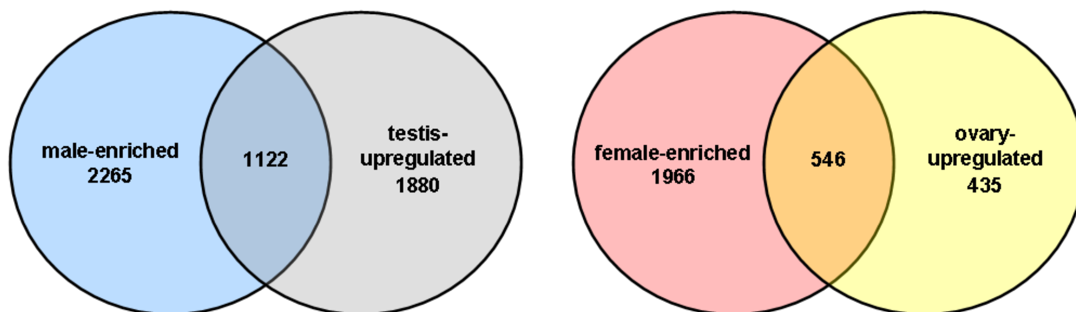

Figure 2 Overlap of Sex- and Gonad-biased Gene Expression.
Male (left) and female (right) Venn diagrams, demonstrating the proportion of genes that fall into both sex- and gonad-biased expression categories. These numbers are based on a "strict consensus" FDR = 0.05, and no fold change threshold. Roughly 33% of male-enriched genes are also significantly testis-upregulated, whereas approximately 22% of female-enriched genes are also significantly ovary-upregulated.

validation experiment is shown in Table 7, and raw expression values and statistical tests are reported in Additional File 6 (Appendix). We calculated 95% confidence intervals for transcript abundance, and hence fold change, (N = 4 - 5) in each case. The confidence intervals are extremely wide for many of the genes, probably due to real variation among individual fish and a small sample size. Nevertheless, each independent test confirmed a significant expression bias in the expected direction, and confidence interval width seems to scale with variation in array fold change estimates across the four different absolute expression analyses.

Table 7 Real-time PCR Confirmation for Seven Differentially Expressed Genes. Expression levels of sex- and gonad-biased zebrafish genes, as confirmed by quantitative real-time PCR. Included are each gene's expression bias category, GenBank identifier and accession number, within-category expression rank, four microarray fold change estimates based on different absolute expression analyses, and qPCR 95% confidence interval for fold change. Some of the confidence intervals are quite wide, but in every case statistically significant ($p < 0.05$) expression bias was confirmed.

Gene class	Gene Name/ GenBank acc. #	Rank	GCOS fold	GC-RMA fold	PM fold	PMMM fold	qPCR 95% CI
Ovary-upregulated	*casp3a* NM_131887	121	3.0	4.7	2.4	2.7	4.6 -15
Ovary-upregulated	zgc:92067 NM_001002377	187	2.4	6.1	2.5	2.7	13 – 57
Testis-upregulated	*sept4* NM_001082815	1	590	1162	47	161	189 – 518
Testis-upregulated	zgc:92129 NM_001002357	10	349	1785	69	61	740 – 3665
Male-enriched	fx05c05.x1 BM571726	4	810	545	38	283	189 – 488
Female-enriched	wu:fd20g04 XM_001334198	1	458	2190	246	551	1982 - >9999
Female-enriched	wu:fd14c01 XM_677844	2	572	1458	174	479	609 – 3246

Discussion

Masculinization of the zebrafish transcriptome

Our results are consistent with the predictions of the "male sex drive" hypothesis. Three lines of evidence from our study provide reason to believe that gene expression in the zebrafish lineage is "masculinized." First, we discovered a larger total number of male-enriched than female-enriched genes (Table 3), consistent with other animal studies. A recent study, for example, documented this asymmetry in five *Drosophila* species (*D. melanogaster*, *D. simulans*, *D. yakuba*, *D. ananasse*, and *D. virilis*) using species-specific microarrays (Zhang et al. 2007), and additional investigations have reported similar findings in *Drosophila* (Parisi et al. 2003; Parisi et al. 2004; Singh and Kulathinal 2005). Rin et al. also identified a substantially greater number of male-enriched genes, especially within higher fold change classes, based on a transcriptomic comparison of testis and ovary in mice (Rinn et al. 2004). In two closely related frog species (*Xenopus laevis* and *X. muelleri*), Malone et al. revealed a greater overall number of male-enriched genes and demonstrated an even more pronounced male-biased asymmetry among genes that are also differentially expressed between species (Malone et al. 2006). Indeed, others have described a related phenomenon, in which male-enriched genes are greatly overrepresented among groups of genes that demonstrate intra- and inter-specific expression polymorphism, relative to female-enriched and sex-unbiased loci (Jin et al. 2001; Meiklejohn et al. 2003; Ranz et al. 2003).

Interestingly, recent studies of sex-biased gene expression in *Danio rerio* have not yielded the same observation of more male-enriched than female-enriched

transcripts. In fact, Santos et al. compared ovary and testis transcriptomes in adult zebrafish and reported 1370 male-enriched genes and 1570 female-enriched genes (Santos et al. 2007), which contrasts with our finding that more genes are male-enriched. One possible source of the discrepancy might be that the experimental animals were treated quite differently in our study. Santos et al. sampled individuals from a "breeding colony" of six males and six females, and histological analysis of experimental ovaries revealed great variation in oogenic stage among individual females (Santos et al. 2007). Females in our study spawned on the same day, and were then isolated from males for five days before being sacrificed. Separation of males and females may not reflect conditions zebrafish experience in nature, but our design allowed us to prevent re-mating and standardize reproductive cycles among experimental individuals. Still, a five-day absence of any stimuli produced by the opposite sex might result in significant behavioral and physiological consequences for males and females, and these could explain the differences between the studies. For example, significant changes in gene expression over a very short time period as a consequence of courtship exposure have been documented in *Drosophila* (Carney 2007). Additional studies should be conducted to assess the potential for plasticity of sex biases in the transcriptome due to behavioral, environmental, developmental (Arbeitman et al. 2002; Sadate-Ngatchou et al. 2004), and temporal factors.

Differences in array platform and analysis might also explain the discrepancy between studies. Santos et al. (Santos et al. 2007) employed microarrays constructed from the Sigma-Genosys (Cambridge, UK) Zebrafish OligoLibrary™, which represents

approximately the same number of unique transcripts (15,806) as the Affymetrix arrays (14,900), but not necessarily the same transcripts. Furthermore, the expression detection algorithms tailored for Affymetrix GeneChips® are unique, and we applied four of these in this study. It is also worth noting that the microarray fold change estimates from the Santos et al. study are substantially lower (up to two orders of magnitude) than the corresponding real-time qPCR fold change estimates, which the authors attribute to spot saturation (Santos et al. 2007). Our microarray fold change estimates appear to be more consistent with the real-time qPCR estimates (Table 7), suggesting that array feature saturation is less of a problem in our study. Despite the discrepancy, however, there is agreement between the two studies at the level of expression patterns for individual genes, as nine out of ten top sex-biased genes identified by Santos et al. (Santos et al. 2007) also appear in our sex-biased gene list (Table 4).

Two other studies addressed sex-biased gene expression in zebrafish, but neither of them is as relevant to this study as the Santos et al. experiment. Wen et al. conducted a whole body male-female comparison of the zebrafish transcriptome using a cDNA microarray representing 8793 unique EST clusters (Wen et al. 2005). The authors identified 383 female-enriched genes in their study; however, they make no mention of male-enriched transcripts, and gonads were not analyzed separately. Another microarray study, by Sreenivasan et al., did separate the gonads, in addition to the brain and kidney, from the "rest-of-body," for males and females (Sreenivasan et al. 2008). They employed cDNA microarrays containing 6370 unique genes derived from zebrafish gonad EST libraries. Sreenivasan et al. reported 881 genes enriched by ≥ 1.5 fold in the

testis relative to the common reference control, and 1366 genes enriched by ≥ 1.5 fold in the ovary relative to the common reference control (Sreenivasan et al. 2008). The report does not provide details regarding the total numbers of male- and female-enriched genes for each organ comparison, so a direct comparison between this study and ours is difficult.

Another surprising result is that we did not identify genes that, according to our strict consensus criteria, demonstrate sex-biased expression at the level of the zebrafish body. A recent study of sex differences with respect to hepatic gene expression, which also utilized the Affymetrix platform, revealed 1249 sex-biased genes (792 male-enriched, 650 female-enriched) in the adult zebrafish liver (Robison et al. 2008). Another study, which examined sex differences of the zebrafish brain transcriptome, identified 42 sex-biased genes (18 male-enriched, 24 female-enriched) (Santos et al. 2008). This is in stark contrast to Sreenivasan et al. (Sreenivasan et al. 2008), who report 3080 genes as differentially expressed between male and female brains, so it is clear that major differences exist among the other zebrafish studies as well. Our study did not involve a direct organ-to-organ comparison (except for gonads), so it is possible that organ-specific signals of sex-biased gene expression were obscured by background gene expression in other somatic tissues. The lack of sexually dimorphic body gene expression in our study could also be a consequence of high among-individual variance in body gene expression, although we took many steps experimentally to reduce this. Furthermore, our statistical criteria for differentially expressed genes were very conservative, so we likely missed some differentially expressed genes, especially if the

differences were small. If we relax our criteria and consider a gene differentially expressed if it appears significant in at least one of the four absolute expression comparisons, then we find 112 body sex-biased genes (78 male-enriched, 34 female-enriched). Of these genes, 26 (9 male-enriched, 17 female-enriched) were consistent with the liver results from Robison et al. (Robison et al. 2008), but none were consistent with the brain study (Santos et al. 2008). The list of 112 genes, and corresponding fold change estimates from the four absolute expression comparisons are included as Additional File 3 (Appendix).

The second pattern indicative of a masculinized transcriptome is an increase in the magnitude of differential expression (i.e. fold change) for male-enriched genes relative to female-enriched genes. Based on our results in *Danio*, male-enriched genes on average demonstrate more extreme sex-biases in expression than female-enriched genes (Figure 1). This trend was also described by Zhang et al. across seven different *Drosophila* species (Zhang et al. 2007). If transcript abundance is viewed as a quantitative trait, it becomes apparent that males demonstrate considerably more exaggerated trait values for sex-biased genes than do females. In essence, for traits that are sexually dimorphic (i.e. expression levels of sex-biased genes), males on average appear to demonstrate more extreme phenotypes. This concept should be relevant to an integrated understanding of transcriptomic masculinization, "male-driven" evolution, and sexual dimorphism at additional phenotypic levels.

A third result of our study related to reproductive processes and sex-specific gene expression patterns is simply that adult male zebrafish demonstrated many more gonad-

soma differences in transcript abundance than females. We detected 5340 genes as differentially expressed between testicular and male body tissue (3002 testis-upregulated, 2338 testis-downregulated). In comparison, only 2380 genes were identified as being differentially expressed between ovarian and female body tissue (981 ovary-upregulated, 1399 ovary-downregulated). These striking transcriptional differences at a tissue-specific level are likely reflections of fundamental reproductive differences between males and females. A microarray study of *D. melanogaster* adults revealed a similar sex disparity in gonad-biased gene expression and also reported that the expression magnitude of testis-upregulated genes is substantially greater than that for ovary-upregulated genes (Parisi et al. 2004). Because none of the 981 ovary-upregulated genes identified in our study demonstrated fold change values greater than four, whereas fold change values for 554 testis-enriched genes exceeded six, zebrafish may also conform to this pattern. A general interpretation of this trend might be that there are more specific transcripts essential to processes that take place in the testes, relative to specific transcripts in ovarian tissue.

A small comparison of testis-upregulated or testis-specific genes from other zebrafish studies (Li et al. 2004; Sreenivasan et al. 2008) to those identified as testis-upregulated in our study indicates a high level of agreement (see "testis-upregulated" section of Table 4). In contrast, many of the top ovary-specific or ovary-upregulated genes identified consistently in these studies are absent from our list of top ovary-upregulated genes (Table 6). Why our study differs from the others in this respect remains an open question. Again, the fact that we separated males from females five

days prior to sample collection may partially explain the discrepancy, especially if females experience major changes in hormone profiles in the absence of males. High body gene expression variance among females in our samples could also explain why ovary-upregulated genes from the other studies did not demonstrate statistically different expression levels in our study. Additional File 7 (Appendix), a more detailed version of Table 3, includes ten reportedly ovary-upregulated genes and the relevant expression value means, standard errors, and fold change estimates from our data set.

A particularly important class of female reproductive genes, which correspond to members of the zona pellucida egg coat glycoprotein superfamily, demonstrate ovary-specific expression patterns according to several zebrafish studies (*zp1* (Zeng and Gong 2002); *zp2* (Wang and Gong 1999; Zeng and Gong 2002); *zp3* (Wang and Gong 1999; Del Giacco et al. 2000)). We, however, identified none of the zona pellucida homologs represented on the zebrafish GeneChip® as significantly ovary-upregulated (See Additional File 8 (Appendix) for a list of *zp* genes, expression value means, and standard errors for each absolute expression analysis). This result is surprising, and the expression values in Additional File 8 indicate high female body *zp* expression in addition to expectedly strong expression in ovaries. Contamination of the body sample with ovarian tissue could produce this result but is unlikely since we completely removed all visible ovarian tissue from each individual. Even if a dissection left as much as half of the total ovarian tissue inside a body sample, one would not expect equal or greater body transcript abundance (for a truly ovary-upregulated gene), because the contaminating ovary signal would be greatly diluted by the female body RNA. Furthermore, if the

female body samples were contaminated with ovarian tissue, we would expect many false positives with respect to male and female body differences, which is clearly not the case. We, therefore, maintain that high female body *zp* expression in our experiment is either real or a reflection of problematic *zp* array probesets. In general, there seems to be some disagreement across studies with respect to tissue specific patterns of *zp* gene expression. For example, significant expression of *zp1* and *zp2* has been documented in ovary-excised females (Wen et al. 2005), and expression of *zp3* in female skeletal muscle has also been described (Zeng and Gong 2002). Furthermore, a recent study (which also used Affymetrix zebrafish arrays) of sex-biased gene expression in the liver of zebrafish reported that *zp2.2*, *zp3*, *zp3a.1*, *zp3b*, and *zpcx* are all expressed at high levels and are all female-enriched (Robison et al. 2008). Based on an estimate by Liu et al., there are likely 10 - 15 *zp2* and 17 - 21 *zp3* paralogs alone distributed throughout the zebrafish genome (Liu et al. 2006), so assaying expression of individual paralogs may not be as straightforward as is assumed. We cannot say for certain that our results reflect this specific problem, but across-study differences in *zp* probe composition might explain some of the inconsistencies in tissue-specific expression patterns of zona pellucida genes.

Genomic differences and sex-biased gene expression

In the absence of dosage compensation, having two copies of a sex chromosome (i.e. the homogametic sex) could allow increased expression of sex chromosome genes in the homogametic sex relative to the heterogametic sex (Ellegren et al. 2007). This is not

likely the reason for sex-biased gene expression in zebrafish, however, because karyotypes of the *Danio rerio* genome fail to reveal heteromorphic sex chromosomes (Traut and Winking 2001). Furthermore, no sex-linked genetic markers or key sex-determination loci have been described in zebrafish as of the completion of our study (von Hofsten and Olsson 2005; Streelman et al. 2007). This suggests that sexually dimorphic gene expression and sexual dimorphism are not explained solely or directly by genome differences between male and female zebrafish. A more plausible scenario is that environmental or genetic conditions initiate sexual differentiation, followed by hormonal differences which cascade into large scale sex-biased gene expression and ultimately into other phenotypic aspects of sexual dimorphism, such as morphological and behavioral differences.

The evolution of sex-biased gene expression

Our study does not specifically address mechanisms potentially responsible for the adaptive evolution of sexually dimorphic gene expression, but these are worth considering here briefly. In general, two processes are capable of generating selection for differential transcript abundance in males and females. Sexual selection could drive the evolution of transcript abundance via mating or fertilization advantages to individuals within a population. Because the general intensity of sexual selection may be different between the sexes (Bateman 1948), it could generate an antecedent for different adaptive trajectories between males and females. Similarly, sex-specific ecological selection could drive the evolution of gene expression via survival, fecundity, or fertility

advantages to members of one or the other sex. If there is intrinsic sex-limitation of the novel transcript abundance from the outset, owing to existing sex-differences in genetic background for example, sexual selection or sex-specific ecological selection can automatically result in sexual dimorphism. If not, a secondary mechanism such as intersexual conflict is required to reinforce stable sexual dimorphism in transcript abundance. Under this scenario, a transition to the male- or female- selected expression "optimum" is constrained, due to a different optimum in the opposite sex. This process generates selection for sex-limited gene expression, and sexually dimorphic expression is a possible response.

Few attempts have been made to rigorously test which (if any) of these processes are responsible for the great degree of sex-biased gene expression observed across animal taxa, but work by Connallon and Knowles (Connallon and Knowles 2005) suggests a signature of sexual conflict in *Drosophila* sex-biased gene expression patterns. Sexual selection in zebrafish has not been quantified formally, but the species exhibits little morphological sexual dimorphism, and observations of mating patterns suggest conditionally high variance in male and female mating success (Spence et al. 2008). More extensive studies comparing gene expression patterns among closely related species that differ with respect to the above selective forces will become feasible in the wake of advancing genomics resources for non-model organisms, and this should greatly improve our evolutionary understanding of sex-biased gene expression.

In general, our microarray results suggest that adult zebrafish demonstrate sexually dimorphic gene expression profiles across a large proportion of the genome. We

detected a greater abundance of male- than female-enriched genes, and found that male-enriched genes demonstrate higher fold changes on average than female-enriched genes. Male zebrafish also demonstrated many more expression differences between body and gonads than did females. These findings are consistent with male-biased patterns of gene expression described in studies of other animal taxa, although they are at odds in some ways with recent zebrafish studies. The discrepancies are discussed, but identifying their sources is difficult due to very different objectives, analyses, and experimental approaches across studies. Sex-biases in gene expression deserve attention because they may explain important differences between males and females, an extension of the realization that gene regulation plays a major role in phenotypic evolution.

3. SEX-, TISSUE-, AND SPECIES-SPECIFIC EXPRESSION PATTERNS INFLUENCE RATES OF SEQUENCE DIVERGENCE BETWEEN TWO HYBRIDIZING SWORDTAIL FISHES

Introduction

Genes expressed differentially between males and females of the same species are of interest to biologists because many of them likely play a role in morphological, physiological, and behavioral differences between the sexes. Phenotypic sexual dimorphism in a species can arise as a result of differing selective regimes acting on males and females, so sex-biased gene expression is of key relevance to those interested in mechanisms of sex-specific selection, namely sexual selection and ecological counterparts such as fecundity selection (Ellegren and Parsch 2007). The last decade of research, fueled initially by microarray technology and now by next-generation sequencing, has supplied many examples of extensive sex-biased gene expression across animals (Jiang et al. 2001; Hahn and Lanzaro 2005; Malone et al. 2006; Zhang et al. 2007; Mank et al. 2008; Small et al. 2009; Prince et al. 2010). In addition to the common occurrence of sex-biased gene expression, molecular evolutionists have shown that protein-coding regions of sex-biased genes, particularly male-biased genes in the *Drosophila melanogaster* subgroup (Meiklejohn et al. 2003; Grath and Parsch 2012), tend to diverge more quickly than genes unbiased with respect to expression in the sexes (Zhang et al. 2004; Cutter and Ward 2005; Mank et al. 2007b).

Although some of the above observations have been interpreted as a pervasive male-specific, transcriptome-wide response to strong sexual selection on male-biased molecules and their expression (Singh and Kulathinal 2005), some caveats are now understood. Female-biased genes, for example, sometimes diverge more rapidly at the amino acid level than male-biased genes, which is the case for autosomal genes expressed in the brain of developing chicken embryos (Mank et al. 2007b). Furthermore, a great deal of sex-biased gene expression in animals is due to major biological differences between male and female reproductive tissues (Rinn and Snyder 2005; Small et al. 2009). Such differences may not reflect recent episodes of sexual selection as much as ancient sex-specific developmental and physiological requirements for producing eggs and sperm.

In general, current observations suggest that sex-biased gene expression and rates of sequence evolution for sex-biased genes are highly context-dependent with respect to the location and timing of transcription. Rapidly-evolving male-biased genes, for example, could be mostly restricted to genes expressed in the male reproductive tract, and male-biased genes in other tissues may not evolve especially rapidly. Furthermore, genes that are expressed with some specificity in certain tissues, regardless of the tissue type, tend to evolve more rapidly than broadly-expressed genes (Duret and Mouchiroud 2000; Zhang and Li 2004), a pattern that also confounds our understanding of sex-biased molecular evolution. These issues have recently been addressed with a re-examination of *Drosophila* microarray and mammalian EST data, in which the author attempts to separate sex- and tissue-specific effects on gene expression and functional sequence

divergence (Meisel 2011). The results from Meisel's study suggest that the relationship between sex-biased gene expression and sequence divergence is driven predominantly by rapidly evolving proteins expressed mostly in the reproductive tract, but similar studies should be executed in groups outside of *Drosophila* and mammals.

We used pyrosequencing to characterize expression biases and coding DNA sequence evolution for gonad, head, and body transcriptomes in male and female *Xiphophorus birchmanni* and *Xiphophorus malinche*. These two hybridizing swordtail fishes, which belong to the live-bearing teleost family Poeciliidae, offer a unique biological backdrop for a study of the relationship between expression patterns and coding sequence divergence. Internal fertilization and postcopulatory influences on mating success are important in other members of Family Poeciliidae (Pilastro et al. 2002; Pilastro et al. 2004), so there is reason to expect sexual selection on male and female reproductive molecules in swordtails. Strong visual and chemical preferences in females with respect to male trait variation have been demonstrated in *X. birchmanni* (Fisher and Rosenthal 2006; Wong and Rosenthal 2006; Fisher and Rosenthal 2007; Willis et al. 2011), so sex- and possibly species-specific patterns of gene expression in "sensory tissues" are predicted. The two species also reside in quite different physical environments regarding elevation, temperature, photic properties, and stream flow dynamics (Rauchenberger et al. 1990; Rosenthal et al. 2003). Adaptive divergence is therefore expected to have generated genomic and phenotypic differences between these two species, a signature of which may be present in large transcript abundance inequalities for many genes.

In this study we analyzed separately the effects of sex-, tissue-, and species-biased gene expression on the functional rate of protein-coding sequence divergence between *X. birchmanni* and *X. malinche* sampled directly from the field. We detected a tremendous amount of sex-biased gene expression across multiple tissue types, along with male and female asymmetries in both the number of sex-biased genes per tissue type and the average magnitude of sex-biased transcript abundance. Sex-biased genes demonstrated especially high d*N*/d*S* ratios relative to unbiased genes for gonad and some non-gonad comparisons. We also found that tissue-biased expression, measured as differences between head and gonad transcript abundance, affects d*N*/d*S* but not likely in an asymmetric fashion. Lastly, we detected a positive relationship between species differences in gene expression and d*N*/d*S*, but only for those comparisons involving the male gonad and the female head transcriptome.

Materials and Methods

Sample preparation and Roche 454 GS FLX sequencing

Eight adult male and eight adult female *Xiphophorus malinche* were captured by baited minnow trap in March 2010 from Chicayotla, Arroyo Xontla in Hidalgo, Mexico. We similarly collected eight adult male and eight adult female *Xiphophorus birchmanni* from the nearby Rio Garces. All females used in this study possessed mature ova but were devoid of developing embryos, in order to minimize variation among individuals and minimize the probability of sampling embryonic tissues. Directly upon taking each fish from the stream, we used sterile dissection tools to remove its gonads (including all

ova in ovaries), major fin tissue (dorsal, caudal, and pectoral), and a dorsal section of the cranium including the sensory organs but excluding the gills. Each tissue and the remainder of each fish's body were immediately frozen individually in TRIzol® Reagent (Invitrogen) aliquots at -80 C until further processing.

In the laboratory we ground the tissues using a Polytron® homogenizer and isolated total RNA from each sample according to the TRIzol® manufacturer's guidelines, except that a "double" extraction was performed by combining the aqueous phase from the initial chloroform separation with a second volume of TRIzol® Reagent and repeating the protocol from the beginning. For each tissue type we pooled 10 µg of total RNA from each of eight individuals, yielding libraries composed of 80 µg of total RNA. The 16 total RNA libraries were as follows: *X. birchmanni* female body, *X. birchmanni* female fins, *X. birchmanni* female gonads, *X. birchmanni* female head, *X. birchmanni* male body, *X. birchmanni* male fins, *X. birchmanni* male gonads, *X. birchmanni* male head, *X. malinche* female body, *X. malinche* female fins, *X. malinche* female gonads, *X. malinche* female head, *X. malinche* male body, *X. malinche* male fins, *X. malinche* male gonads, and *X. malinche* male head. We sent all pooled samples to the Michigan State University Research Technology Support Facility, where Rapid Prep cDNA libraries were generated, multiplexed, and sequenced in two runs on a Roche 454 GS FLX® sequencer using Titanium® chemistry.

De novo transcriptome assembly, alignment of orthologous coding sequences, and transcript abundance estimates

We trimmed all 454 reads to remove low quality regions, polyA tails, and remaining library preparation artifacts using the program SeqTrim (Falgueras et al. 2010). After discarding all cleaned reads less than 50 nt, we performed two *de novo* transcriptome assemblies (one for each species) using the CLC Genomics Workbench® version 4.5 (CLC bio). Aside from default assembly parameters, we selected the remapping option with similarity criterion set at 0.97 for the final assemblies. Nucleotides in the final contig sequences were determined by a majority consensus, that is, the most common base among all assembled reads for a given position.

Between-species orthologous sequence pairs were identified using a "reciprocal best BLAST hit" criterion (Rivera et al. 1998), wherein the BLAST hit (Altschul et al. 1990) for each search with the highest bit-score was used to establish the "best hit." This approach mandates that orthologs are only identified when two sequences from different species are each other's top BLAST hit. After obtaining a set of putatively orthologous transcripts for the two species, we used BLASTx (Camacho et al. 2009) to obtain homologous protein-coding references from the NCBI non-redundant protein sequence database. We translated *X. birchmanni* and *X. malinche* high-scoring segment pairs (HSPs) from the BLASTx output into amino acids to define open reading frames, aligned them using Clustal Omega (Sievers et al. 2011), and then reverse-translated the alignment after excluding error-prone regions with a custom sliding window script (by

R. Cui). A minimum alignment length of 70 codons was set to filter especially small transcript fragments from downstream analyses.

For each orthologous contig containing an open reading frame greater than 70 codons we estimated transcript abundance in 12 libraries of interest (*X. birchmanni* female body, *X. birchmanni* female gonads, *X. birchmanni* female head, *X. birchmanni* male body, *X. birchmanni* male gonads, *X. birchmanni* male head, *X. malinche* female body, *X. malinche* female gonads, *X. malinche* female head, *X. malinche* male body, *X. malinche* male gonads, and *X. malinche* male head) by mapping sequencing reads from a given library back to its respective species' assembly with the RNA-seq module in the CLC Genomics Workbench®. Transcript abundance was expressed as *RPKM,* the number of reads per kilobase of contig per one million mapped reads (Mortazavi et al. 2008). All expression ratios for genes calculated in this study are simply *RPKM* quotients.

Molecular evolutionary and statistical analyses

A pairwise maximum likelihood estimate of d*N*/d*S*, the ratio of the nonsynonymous substitution rate to the synonymous substitution rate, was obtained from each aligned pair of protein-coding sequences using the codeml program (runmode = -2) within PAML 4.5 (Yang 1997). Higher d*N*/d*S* values for particular genes or lineages suggest a greater extent of diversifying selection or a relative relaxation of purifying selection (Yang 1998). Because the interpretation of d*N*/d*S* is questionable when the denominator is zero, all observations of this nature were excluded from analysis. All statistical tests

were either performed in Microsoft Excel® (Microsoft Corp.) or JMP Pro® version 9 (SAS Institute Inc.). We analyzed the relationship between d*N*/d*S* and expression differences separately on three levels: sex-biased gene expression, tissue-biased gene expression, and species-biased gene expression.

Initially we assessed the general direction and magnitude of male-and female-biased transcript abundance for six male-female library pairs (*X. birchmanni* body, *X. birchmanni* gonads, *X. birchmanni* head, *X. malinche* body, *X. malinche* gonads, and *X. malinche* head). For each transcript we calculated the degree of sex-biased gene expression as

$$\log \frac{RPKM_{male_{ij}}}{RPKM_{female_{ij}}},$$

where i is one of the six library types listed above and j represents the individual transcript being assessed. To coarsely categorize transcripts as male- or female-biased, we used the equivalent of an approximately two-fold expression difference cutoff, such that if

$$\frac{RPKM_{male_{ij}}}{RPKM_{female_{ij}}} > 2,$$

the gene was considered male-biased, and if

$$\frac{RPKM_{male_{ij}}}{RPKM_{female_{ij}}} < 0.5,$$

the gene was considered female-biased. Genes with an *RPKM* value of zero in both sexes for a given library comparison were excluded from analysis. For each of the six library types we calculated the relative number of male- and female-biased genes using

these criteria, and then we compared d*N*/d*S* among male-biased, female-biased, and unbiased categories. A second "continuous" analysis, considering effects across all six library types simultaneously, was carried out to examine the nature of the relationship between d*N*/d*S* and the magnitude of sex-bias without regard to the direction of bias. To accomplish this we fit a generalized linear model to the data, wherein the distribution of the response variable d*N*/d*S* was modeled as exponential with an inverse link function. The six explanatory terms in the model are expressed as

$$\left|\log \frac{RPKM_{male_{ij}}}{RPKM_{female_{ij}}}\right|.$$

The relationship between d*N*/d*S* and tissue specificity with respect to head and gonad library pairs was assessed in a similar manner. For each transcript we calculated the degree of tissue-biased gene expression as

$$\log \frac{RPKM_{gonad_{ij}}}{RPKM_{head_{ij}}},$$

where *i* is one of four gonad-head library pairs (*X. birchmanni* male, *X. birchmanni* female, *X. malinche* male, and *X. malinche* female), and *j* represents the individual transcript being assessed. Using the same strategy described for the sex-biased analysis, we compared d*N*/d*S* among gonad-biased, head-biased, and unbiased groups of genes separately for *X. birchmanni* male, *X. birchmanni* female, *X. malinche* male, and *X. malinche* female library pairs. Likewise, we fit a generalized linear model to explain variation in d*N*/d*S* as a function of four effect terms, which may be expressed as

$$\left|\log \frac{RPKM_{gonad_{ij}}}{RPKM_{head_{ij}}}\right|,$$

where i represents each of the four gonad-head library pairs and j represents each individual transcript.

Finally, to explore the relationship of protein-coding sequence divergence to expression divergence between species, we calculated the degree of species-biased gene expression as

$$\log \frac{RPKM_{X.birchmanni_{ij}}}{RPKM_{X.malinche_{ij}}},$$

where i is one of six *X. birchmanni*-*X. malinche* library pairs (male body, male gonad, male head, female body, female gonad, and female head), and j represents the individual transcript being assessed. Again, we fit the same type of generalized linear model with the six effect terms expressed as

$$\left|\log \frac{RPKM_{X.birchmanni_{ij}}}{RPKM_{X.malinche_{ij}}}\right|,$$

where i represents each of the four library pairs specified above and j represents each individual transcript.

Results

454 sequencing, assembly, and orthology assignment

The two sequencing runs yielded a total of 2,136,022 passing reads with a mean read length of 306 nt. The number and mean length of reads for each library are reported in Table 8, along with assembly results for the two species. An especially poor yield was obtained for the *X. malinche* female fin sequencing library, so we excluded all fin libraries from analysis in this study. Also, the *X. malinche* female body library consisted

of many fewer reads than other libraries, so expression ratio results involving this library should be interpreted cautiously. The two *de novo* assemblies resulted in a mean contig number of 57,559 and a mean contig length of 684 nt, on par with or better than 454 *de novo* transcriptome assemblies for a congener (Zhang et al. 2011) and other teleosts (Elmer et al. 2010). We identified 31,991 putative orthologous pairs, based on our reciprocal best BLAST criteria. Of these, however, only 10,222 were retained for downstream analysis. The remaining 21,796 orthologous pairs were either devoid of an open reading frame (most BLASTed to 3' or 5' UTRs of known genes) or consisted of an amino acid alignment of less than 70 residues.

Sex-biased gene expression and dN/dS

All three tissue types (body, gonad, and head) in both species demonstrated a significant amount of sex-biased gene expression. On average, ~71% of the transcripts represented in a given library type were sex-biased, based on our two-fold-difference criterion. We detected a strong asymmetry in the proportion of male- versus female-biased genes in all tissue types (Table 9). There were significantly more male-biased than female-biased genes in body and head library types for both species, but significantly more female-biased than male-biased genes in gonad library types for both species ($p < 0.0001$, sequential *G*-tests for goodness of fit). Multiple Mann-Whitney U tests also suggested that the *magnitude* of sex-biased expression adheres to this same pattern, with the exception of no significant difference between male and female *X. birchmanni* body libraries (Table 9).

We also detected differences in d*N*/d*S* among sex-biased and unbiased transcripts for several library types (Figure 3). Both male- and female-biased genes expressed in *X. birchmanni* gonads demonstrated a higher d*N*/d*S* than *X. birchmanni* gonad genes unbiased with respect to sex (Kruskal-Wallis Test, $p = 0.0002$). The same trend was observed for *X. malinche* body-expressed genes (Kruskal-Wallis Test, $p = 0.0050$) and *X. malinche* gonad genes (Kruskal-Wallis Test, $p = 0.0156$). *X. malinche* sex-biased genes (especially female-biased genes) expressed in the head also tended towards having

Table 8 *Xiphophorus* 454 Sequencing and *de novo* Transcriptome Assemblies.

<table>
<tr><th>Sequencing Library</th><th>Read #</th><th>Read Length Mean (nt)</th><th>Contig #</th><th>Contig Length Mean (nt)</th></tr>
<tr><td>X. birchmanni ♂ Body</td><td>126,735</td><td>306</td><td rowspan="8">57,063</td><td rowspan="8">657</td></tr>
<tr><td>X. birchmanni ♂ Fins</td><td>158,543</td><td>302</td></tr>
<tr><td>X. birchmanni ♂ Gonads</td><td>142,000</td><td>327</td></tr>
<tr><td>X. birchmanni ♂ Head</td><td>131,422</td><td>310</td></tr>
<tr><td>X. birchmanni ♀ Body</td><td>86,289</td><td>282</td></tr>
<tr><td>X. birchmanni ♀ Fins</td><td>88,729</td><td>303</td></tr>
<tr><td>X. birchmanni ♀ Gonads</td><td>182,941</td><td>315</td></tr>
<tr><td>X. birchmanni ♀ Head</td><td>116,629</td><td>284</td></tr>
<tr><td>X. malinche ♂ Body</td><td>216,295</td><td>329</td><td rowspan="8">58,054</td><td rowspan="8">710</td></tr>
<tr><td>X. malinche ♂ Fins</td><td>205,952</td><td>324</td></tr>
<tr><td>X. malinche ♂ Gonads</td><td>145,087</td><td>325</td></tr>
<tr><td>X. malinche ♂ Head</td><td>170,473</td><td>328</td></tr>
<tr><td>X. malinche ♀ Body</td><td>39,182</td><td>280</td></tr>
<tr><td>X. malinche ♀ Fins</td><td>2,781</td><td>223</td></tr>
<tr><td>X. malinche ♀ Gonads</td><td>173,162</td><td>322</td></tr>
<tr><td>X. malinche ♀ Head</td><td>149,802</td><td>317</td></tr>
</table>

higher d*N*/d*S* ratios relative to unbiased genes (Kruskal-Wallis Test, $p = 0.0097$).

A generalized linear model for variation in d*N*/d*S* as a function of sex-biased gene expression across all six library types identified significant effects from *X. birchmanni* gonad, *X. malinche* body, and *X. malinche* head library pairs. Results from the model are reported in Table 10.

Table 9 Sex-biased Gene Expression Trends in *Xiphophorus.*
Sex-asymmetries in the number of sex-biased genes and the magnitude of sex-biased expression for each tissue type. * The *X. malinche* female body library contains an especially low number of sequencing reads, making comparisons less robust.

Tissue (Species)	# of Male-biased Transcripts	# of Female-biased Transcripts	Sex Asymmetry in Number of Sex-biased Transcripts?	Sex Asymmetry in Magnitude of Sex-biased Expression?
Body *X. birchmanni*	3192	1446	Yes, M > F (p < 0.0001)	No (p = 0.2765)
Gonad *X. birchmanni*	2452	3627	Yes, F > M (p < 0.0001)	Yes, F > M (p = 0.0046)
Head *X. birchmanni*	3093	2160	Yes, M > F (p < 0.0001)	Yes, M > F (p < 0.0001)
Body *X. malinche* *	5472	931	Yes, M > F (p < 0.0001)	Yes, M > F (p < 0.0001)
Gonad *X. malinche*	2426	3325	Yes, F > M (p < 0.0001)	Yes, F > M (p < 0.0001)
Head *X. malinche*	2539	2193	Yes, M > F (p < 0.0001)	Yes, M > F (p < 0.0001)

Tissue-biased gene expression and dN/dS

Categorical comparisons of *dN/dS* among gonad-biased, head-biased, and unbiased groups of genes revealed a positive relationship between tissue specificity in general and *dN/dS* (Figure 4). Both ovary-biased and head-biased transcripts in *X. birchmanni* females demonstrated a higher *dN/dS* than genes expressed evenly between the two

Table 10 Magnitude of Sex-biased Gene Expression Explains variation in d*N*/d*S*. Generalized linear model for d*N*/d*S* (exponentially distributed with reciprocal link function), as a function of sex-biased gene expression in six different library pairs. Model terms having a significant effect at $\alpha = 0.05$ are in bold. As a result of the reciprocal link function, a negative effect sign should be interpreted as a positive relationship between explanatory and response variables.

Term	ChiSquare	DF	*P*	Effect Parameter Estimate
Model (lnLFull - lnLReduced)	**62.7574**	**6**	**< 0.0001**	
|log(M/F)| *X. birchmanni* Body	0.1050	1	0.7460	0.0669
|log(M/F)| *X. birchmanni* Gonad	**31.0285**	**1**	**< 0.0001**	**-1.1364**
|log(M/F)| *X. birchmanni* Head	0.009216	1	0.9235	0.0203
|log(M/F)| *X. malinche* Body	**19.6744**	**1**	**< 0.0001**	**0.9387**
|log(M/F)| *X. malinche* Gonad	0.4869	1	0.4853	-0.1570
|log(M/F)| *X. malinche* Head	**5.1031**	**1**	**0.0239**	**-0.5935**

tissues (Kruskal-Wallis Test, $p = 0.0006$). We also found differences in *dN/dS* among head-biased, gonad-biased, and unbiased genes in male *X. birchmanni* and male *X. malinche* (Kruskal-Wallis Tests, $p = 0.0146$ and $p = 0.0125$, respectively). The *X. malinche* female comparison, however, suggested minimal variation of *dN/dS* among the three expression groups (Kruskal-Wallis Test, $p = 0.0812$).

A generalized linear model very similar to the one that was fit to the sex-biased expression data identified significant effects of tissue-biased gene expression on *dN/dS*. In particular, we found large positive effects of tissue-biased gene expression in *X. birchmanni* and *X. malinche* females, but a significant negative effect of tissue-biased gene expression on *dN/dS* in *X. malinche* males (Table 11).

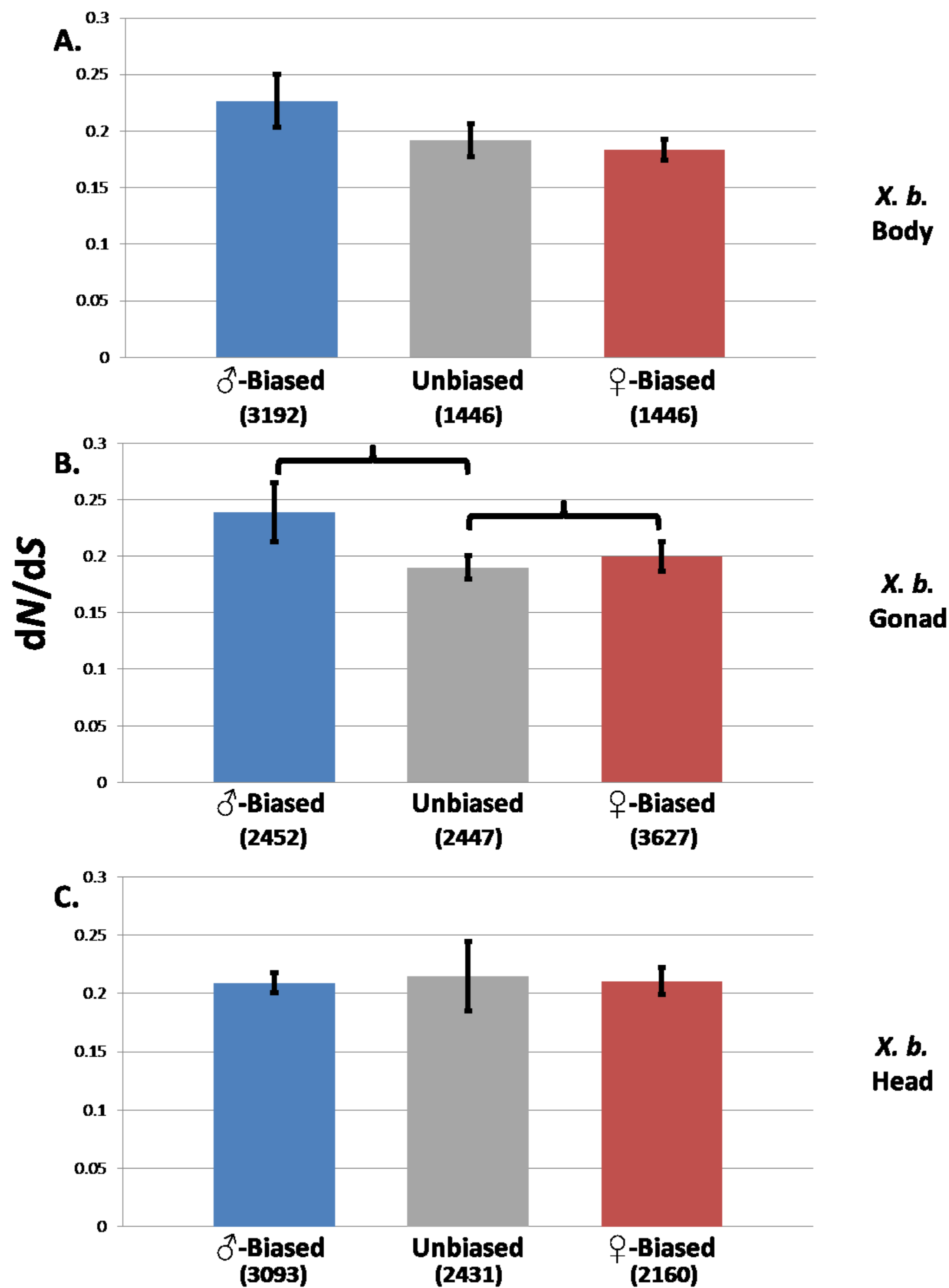

Figure 3 Sex-biased Gene Expression and d*N*/d*S* in *Xiphophorus.*
Mean d*N*/d*S* for male-biased, female-biased, and sex-unbiased genes in all six library types (A. – F.) assessed with respect to sex-biased gene expression. Error bars represent the standard error of the mean. Categories significantly different after three standard Bonferroni-corrected Mann-Whitney U tests are linked by braces.

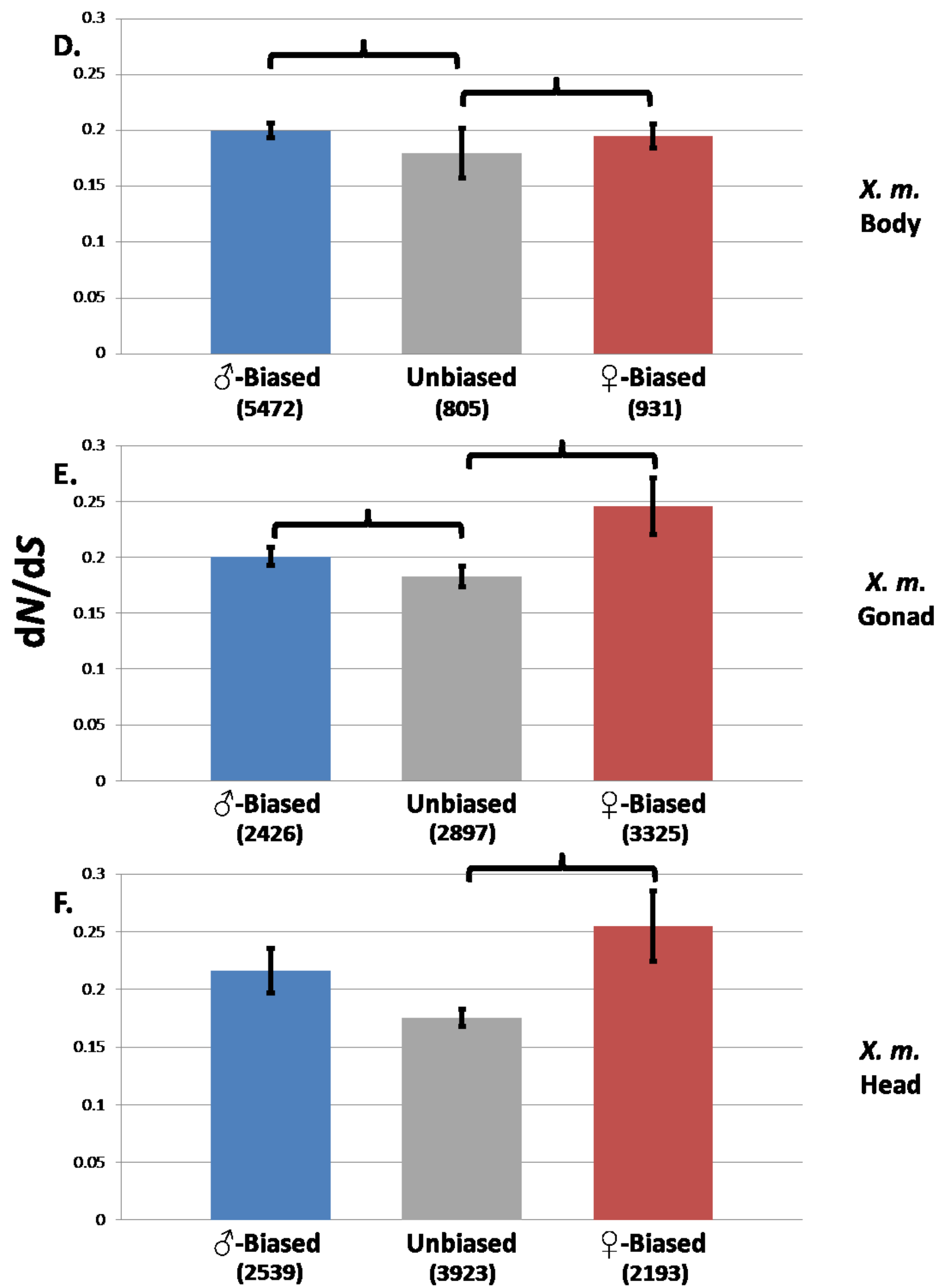

Figure 3 Continued

Table 11 Magnitude of Tissue-biased Gene Expression Explains variation in d*N*/d*S*.
Generalized linear model for d*N*/d*S* (exponentially distributed with reciprocal link function), as a function of tissue-biased gene expression in four different library pairs. Model terms having a significant effect at $\alpha = 0.05$ are in bold. As a result of the reciprocal link function, a negative effect sign should be interpreted as a positive relationship between explanatory and response variables.

Term	ChiSquare	DF	*p*	Effect Parameter Estimate
Model (lnLFull - lnLReduced)	**71.2714**	**4**	**< 0.0001**	
|log(Gonad/Head)| *X. birchmanni* ♀	**10.5696**	**1**	**0.0011**	**-0.4277**
|log(Gonad/Head)| *X. birchmanni* ♂	0.6122	1	0.4339	-0.1036
|log(Gonad/Head)| *X. malinche* ♀	**39.3794**	**1**	**< 0.0001**	**-0.8940**
|log(Gonad/Head)| *X. malinche* ♂	**11.8081**	**1**	**0.0006**	**0.4999**

Expression and sequence divergence between X. birchmanni and X. malinche

We detected a subtle but positive relationship between expression divergence and *dN/dS* with respect to some but not all tissue types examined in this study. The results from our generalized linear model suggest that divergence in transcript abundance between the two species is positively related to sequence divergence when expression comparisons involve female head and male gonad tissues (Table 12; Figure 5 D. and E.). Six scatterplots, demonstrating the nature of the weak relationship between *dN/dS* and each term in the model, are presented in Figure 5.

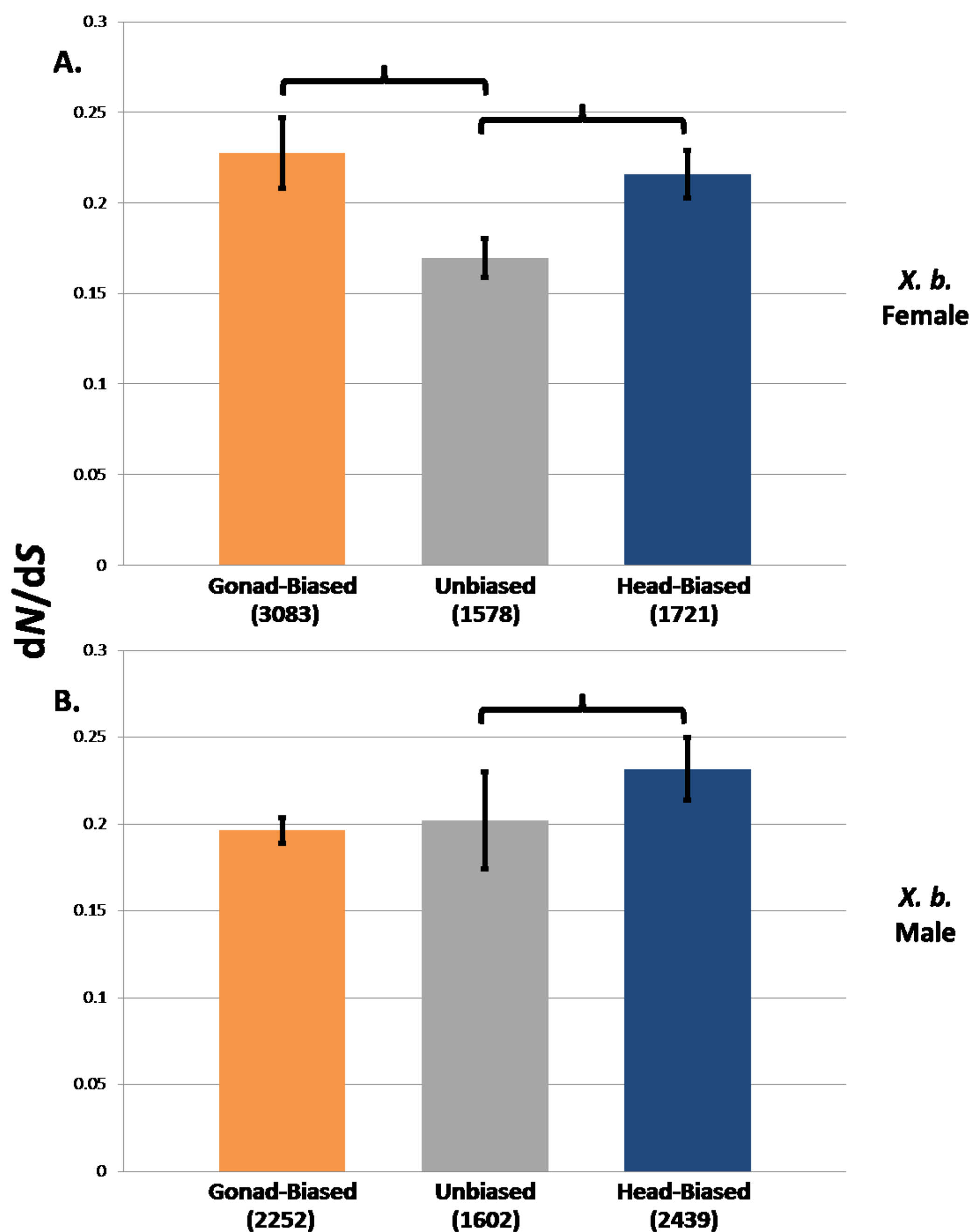

Figure 4 Tissue-biased Gene Expression and d*N*/d*S* in *Xiphophorus.*
Mean d*N*/d*S* for gonad-biased, head-biased, and tissue-unbiased genes in all four library pairs (A. – D.) analyzed with respect to tissue-biased gene expression. Error bars represent the standard error of the mean. Categories significantly different after three standard Bonferroni-corrected Mann-Whitney U tests are linked by braces.

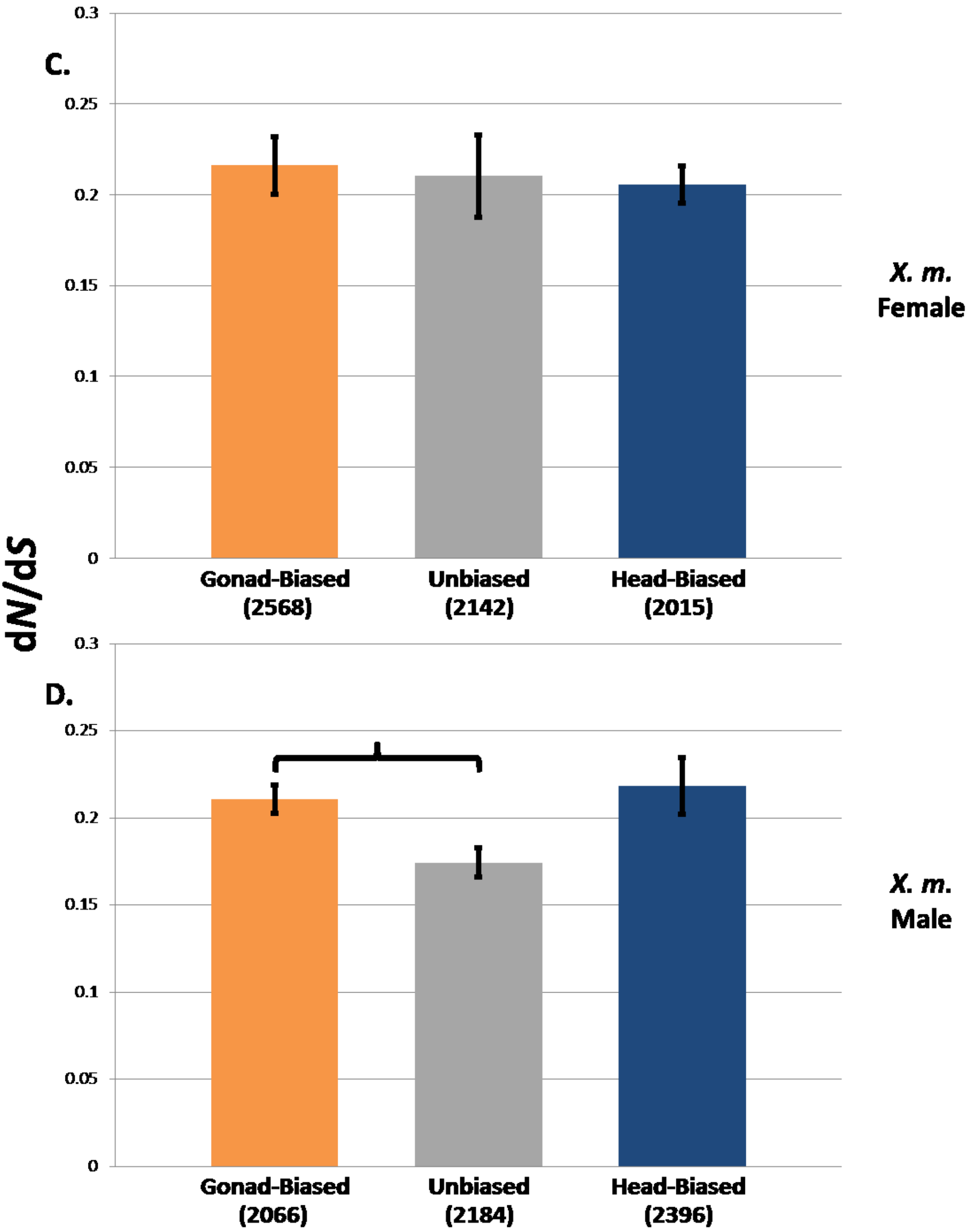

Figure 4 Continued

Table 12 Expression Divergence Between Species Explains Variation in d*N*/d*S*. Generalized linear model for d*N*/d*S* (exponentially distributed with reciprocal link function), as a function of species-biased gene expression in six different library pairs. Model terms having a significant effect at α = 0.05 are in bold. As a result of the reciprocal link function, a negative effect sign should be interpreted as a positive relationship between explanatory and response variables.

Term	**ChiSquare**	**DF**	***P***	**Effect Parameter Estimate**
Model (lnLFull - lnLReduced)	**41.1489**	**6**	**< 0.0001**	
|log(*X. birchmanni*/*X. malinche*)| ♀ Body	3.5074	1	0.0611	0.3927
|log(*X. birchmanni*/*X. malinche*)| ♀ Gonad	0.09781	1	0.7545	-0.09131
|log(*X. birchmanni*/*X. malinche*)| ♀ Head	**19.7800**	**1**	**< 0.0001**	**-1.0507**
|log(*X. birchmanni*/*X. malinche*)| ♂ Body	0.0009553	1	0.9753	0.007805
|log(*X. birchmanni*/*X. malinche*)| ♂Gonad	**8.8954**	**1**	**0.0029**	**-0.7923**
|log(*X. birchmanni*/*X. malinche*)| ♂ Head	0.1816	1	0.6700	0.1244

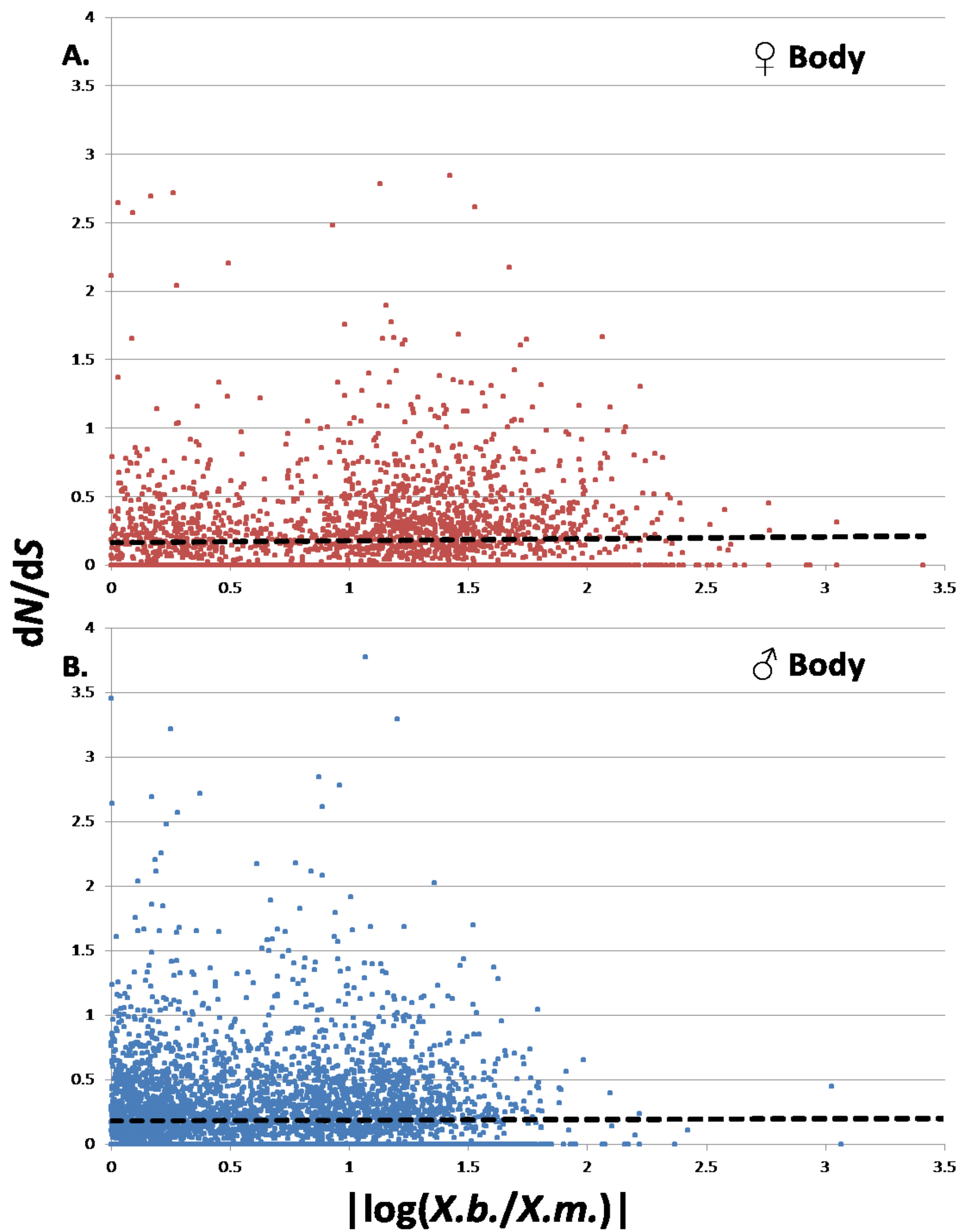

Figure 5 Sequence Divergence as a Function of Expression Divergence. Regression of d*N*/d*S* on the magnitude of expression divergence between *X. birchmanni* and *X. malinche* for all six tissue types (A. – F.). A fitted least-squares regression line (dashed) represents the relative strength of each relationship.

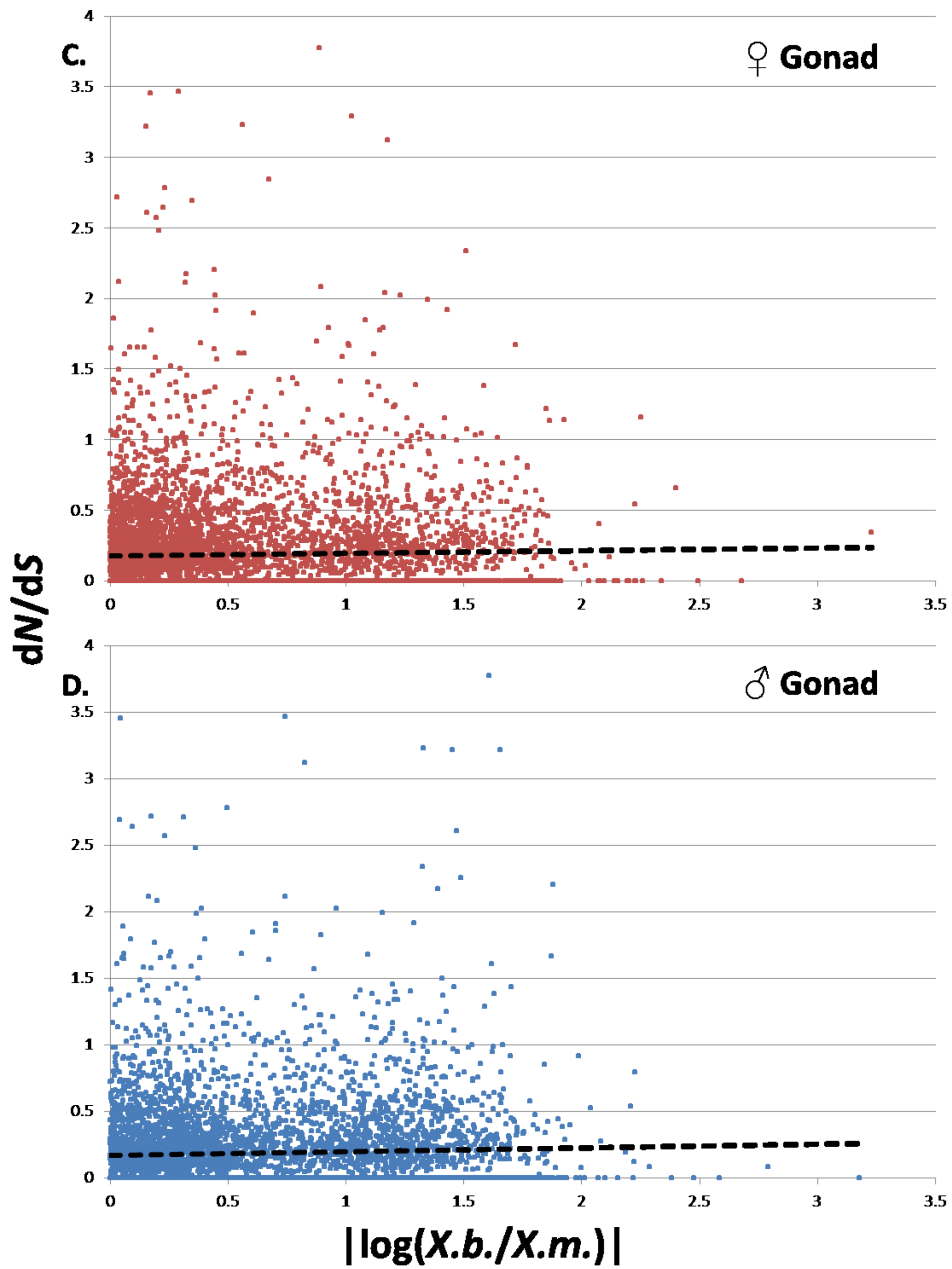

Figure 5 Continued

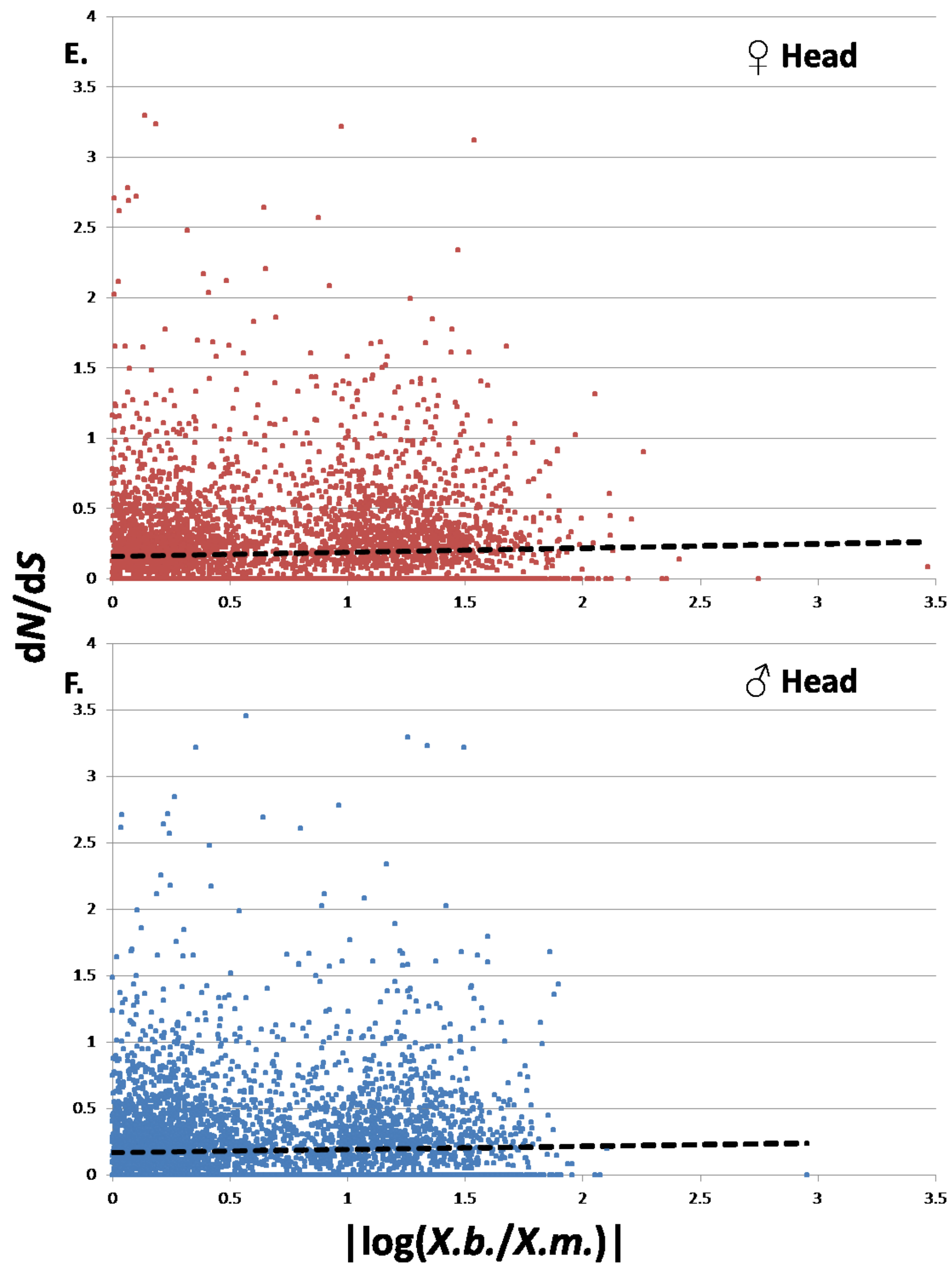

Figure 5 Continued

Discussion

Sex-biased gene expression and coding sequence evolution in Xiphophorus

The fact that such a large proportion of the transcriptome for each of the three tissue types in both species is sex-biased comes perhaps as no surprise, given that 20-40% of the transcriptome may be sexually dimorphic in other animals (Ranz et al. 2003; Rinn et al. 2004; Small et al. 2009). Our observation that nearly 70% (on average) of each transcriptome is sex-biased in *Xiphophorus*, however, is almost certainly inflated due to our relatively low throughput, non-replicated assays of transcript abundance. The fact that the *X. malinche* body library consisted of relatively few reads is particularly problematic, as sampling bias reduces the reliability of any expression comparison involving this library. The results of this problem are apparent in the male-female *X. malinche* body comparison (Figure 4 D.), which reflects an overestimation of the number of male-biased genes.

We found a general asymmetry in the number of male- versus female-biased genes for each of the three tissue types, which was concordant between the two species (Table 9). For head and body tissue types we observed a greater number of male-biased genes, a pattern commonly reported among vertebrates for testis-ovary comparisons (Rinn et al. 2004; Malone et al. 2006; Small et al. 2009) and for most *Drosophila* species in general (Zhang et al. 2007). Interestingly, we found a greater number of female-biased genes for the *Xiphophorus* gonad comparison, which is a pattern consistent with male-female comparisons of somatic tissues in other animals (Ranz et al. 2003; Yang et al. 2006). The average magnitude of expression bias for female-biased genes in

Xiphophorus gonads was also greater than that for male-biased gonad genes (Table 9), which is also at odds with results from zebrafish (Small et al. 2009). Why female-biased genes in swordtail gonads surpass male-biased genes in both number and in magnitude of expression bias is unclear. It is possible that one or more of the females used in this study contained early-stage developing embryos, although no embryonic tissue was observed during dissections.

In *X. birchmanni* we found evidence for elevated d*N*/d*S* among sex-biased genes, but only when comparing sexes within the gonad library type (Figure 3 A. – C.). Importantly, there was no difference between d*N*/d*S* among female-biased genes and d*N*/d*S* among male-biased genes, which constitutes a lack of evidence for the kind of uniquely rapid divergence of male-biased genes documented elsewhere (Meiklejohn et al. 2003; Grath and Parsch 2012) . Information from the generalized linear model confirms a general, positive relationship between d*N*/d*S* and the magnitude of within-gonad sex-biased gene expression.

We observed slightly different results concerning sex-biased gene expression in *X. malinche* tissues (Figure 3 D. – F.). As in *X. birchmanni*, genes expressed differentially between ovary and testis have diverged rapidly relative to genes that are not sex-biased in expression, although the generalized linear model failed to detect this effect in continuous terms. Our discrete analysis appears to suggest especially rapid divergence of sex-biased genes in the *X. malinche* body as well, but unreliable expression ratios due to so few reads in female body library render this conclusion dubious at best. According to our statistical model, the term for *X. malinche* within-body

sex bias has a significant but negative effect on d*N*/d*S*. This result is also difficult to interpret given the problematic nature of the female body data for *X. malinche*. Sex-biased genes expressed in the *X. malinche* head, on the other hand, may actually tend to evolve more rapidly, according to our discrete and continuous analyses. Why a parallel relationship does not exist for *X. birchmanni* is puzzling, but any ecological explanation for such a difference would be completely *ad hoc*. At the very least, a well-replicated, higher throughput RNA-seq study might be conducted to confirm the validity of this discrepancy before biological speculation.

Tissue-biased gene expression and coding sequence evolution in Xiphophorus

Because tissue-specificity in general can co-vary with the rate of sequence evolution across the genome, it stands to reason that this explanation should be ruled out before asserting claims, in the absence of data from other tissues, that gonad-specific genes evolve especially rapidly (Duret and Mouchiroud 2000; Meisel 2011). We compared transcript abundances between head and gonad libraries in an attempt to distinguish effects of gonad-specificity from effects of tissue-specificity on d*N*/d*S*. In general our results failed to demonstrate a clean separation of these effects, as we found no evidence to indicate that gonad-biased transcripts evolve more rapidly than head-biased transcripts (Figure 4). Tissue-specificity in general does seem to influence d*N*/d*S* in the expected direction when considering *X. birchmanni* male and *X. birchmanni* female tissues, although a non-significant effect of tissue-biased expression on d*N*/d*S* was inferred from our generalized linear model (Table 11). The results from *X. malinche* are less easily

interpreted. Results from the continuous analysis suggest a relatively strong positive relationship between within-female tissue bias and d*N*/d*S*, but the discrete nonparametric test for a difference in d*N*/d*S* among head-biased, gonad-biased, and unbiased genes is less discerning. The discrete analysis of d*N*/d*S* among expression groups within *X. malinche* males suggests higher d*N*/d*S* values for tissue-biased genes, but results from the generalized linear model imply that d*N*/d*S* is negatively associated with the degree of tissue-biased expression. Despite the ambiguity associated with the *X. malinche* results, we found no strong evidence for particularly rapid evolution of ovary- or testis-biased genes relative to head-biased genes. Given that gonad-specific genes do evolve more rapidly than other tissue-specific genes in *Drosophila* and mammals (Meisel 2011), our results in *Xiphophorus* either reflect a true paucity of rapidly evolving gonad-specific genes, or demonstrate that head-specific genes as a group have also diverged rapidly. Data from other specific tissues such as liver, spleen, and kidney would shed light on this particular issue.

Expression and sequence divergence between X. birchmanni *and* X. malinche

A positive, genome-wide relationship between sequence divergence and divergence in transcript abundance has been documented in *Drosophila* (Nuzhdin et al. 2004; Lemos et al. 2005), but no such relationship was documented in a human-mouse comparison (Jordan et al. 2004). We found some evidence for a weak but positive relationship between these two variables, but only in two out of six tissue types (Table 12, Figure 5 D. – E.). Testes and female head tissues appear to adhere to this pattern, a finding that is

subject to interpretation. Others (Khaitovich et al. 2005) have reported a positive relationship between expression divergence and coding sequence divergence, but only for the testis transcriptome in a human-chimpanzee comparison (brain, heart, kidney, and liver tissues were also studied). In the brain tissue analysis from their study, only male samples were examined, so we cannot say whether the pattern exists for female brains. Expression levels of testis-expressed genes have been shown to diverge especially rapidly (and adaptively) compared with transcriptomes of other tissues in great apes (Khaitovich et al. 2005) and mice (Voolstra et al. 2007).

Sequence and expression divergence between *X. birchmanni* and *X. malinche* may reflect species differences with respect to sperm competition or genetic compatibility, and the divergence is suspected to be relatively rapid owing to sexual selection, sexual conflict, and/or reinforcement of speciation. Our results suggest that these or other selective forces may drive transcript abundance changes and protein divergence simultaneously. According to this logic, however, divergence in expression and sequence evolution are expected to be similarly coupled for ovarian tissue, which is not the case. Also intriguing is the positive relationship between expression divergence and coding sequence evolution for female head tissue. Intuitively, the female brain and sensory tissues should play seminal and possibly unique roles with respect to female mate choice and mate recognition, so perhaps the same evolutionary mechanisms stated above affect testis and female sensory tissues in the same manner.

In general we have demonstrated that multiple “dimensions” of gene expression variation (sex-, tissue-, and species-biased expression) are associated with functional

substitution rates, although each of these relationships is highly context-dependent. Sex-biased genes appear to evolve more rapidly than their unbiased counterparts particularly in the case of gonad-expressed molecules. Tissue-biased genes evolve quickly in many cases regardless of whether they are gonad- or head-biased. Lastly, genes demonstrating species-biased patterns of expression evolve rapidly, but only when assessed within the context of testis and female head transcriptomes. Data collected with better resolution of expression differences in mind, and independent assays of both gene expression and sequence variation among individuals in natural populations of *X. birchmanni* and *X. malinche*, will ultimately be necessary in order to address the precise mechanisms underlying the general relationships presented here.

4. SEXUAL SELECTION AND THE EVOLUTION OF MALE PREGNANCY PROTEINS AMONG PIPEFISH AND SEAHORSE LINEAGES WITH DIVERSE MATING SYSTEMS

Introduction

Reproductive genes are among the fastest-evolving elements of animal and plant genomes (Swanson and Vacquier 2002a; Torgerson et al. 2002). Though multiple causative agents have been proposed, the fundamental reasons for this pattern remain unclear after nearly two decades of research. The most popular hypothesis to date stipulates that postcopulatory sexual selection is the driving force behind the rapid evolution of reproductive molecules, but few studies have rigorously tested this notion on a broad scale. Effective tests of the sexual selection hypothesis should measure the relationship between the strength of sexual selection and the rate of reproductive protein evolution across multiple lineages. If sexual selection is a common driver of rapid reproductive molecular evolution, many orthologs from lineages in which one expects a history of strong postcopulatory sexual selection *a priori* should demonstrate elevated rates of sequence diversification relative to corresponding orthologs from lineages lacking such a history. Given this prediction, an obvious strategy for assessing the generality of the sexual selection hypothesis is to measure lineage-specific rates of substitution for many putative reproductive genes, across a phylogeny of species in which monogamous and polygamous mating systems are represented.

At least six studies, reviewed in (Wong 2011), have applied the above approach for one or several reproductive genes at a time. One study, for example, examined seven male ejaculate proteins using coding sequence alignments from up to 20 rodent species with diverse mating systems (Ramm et al. 2008). The authors used explicit comparisons of codon substitution models in PAML (Yang 1997; Yang 2007) to assess whether these proteins are likely to contain a proportion of sites under diversifying selection specifically in lineages demonstrating high sperm competition. Although five out of the seven ejaculate proteins were in general likely to contain positively selected sites, Ramm et al. discovered a strong lineage-specific effect of mating system for only one of the molecules, a primary component of the copulatory plug. Others (Hurle et al. 2007) compared protein-coding sequences of six physically linked genes with putative reproductive functions across 13 primate species, but failed to detect the predicted relationship between mating system and lineage-specific rates of protein divergence. The evolution of 18 seminal fluid proteins among 13 *Heliconius* butterfly species, representing a pupal-mating monadrous clade and an adult-mating polyandrous clade, has also been characterized recently (Walters and Harrison 2011). Results from this study similarly revealed evidence for a few seminal proteins having diverged as a result of positive selection. However, Walters and Harrison reported higher estimates of protein diversification on average in the monandrous clade, perhaps because monandry is actually the derived mating system, and elevated rates of protein evolution may be due to historically relaxed selective constraints.

Here we took advantage of massively parallel DNA pyrosequencing to survey molecular evolutionary patterns for ~800 genes expressed in the brooding structures of male-pregnant syngnathid fishes. Pipefishes and seahorses provide an ideal taxon in which to test the sexual selection hypothesis for rapid reproductive protein evolution for several key reasons. First, many species within the group are extremely variable and well characterized with respect to mating system and sexual selection, by way of genetic parentage analysis and detailed behavioral studies (Masonjones and Lewis 2000; Jones and Avise 2001; Sogabe and Yanagisawa 2007). Some lineages are rather unique among animals in that total sexual selection is actually stronger on females than on males (Jones and Avise 1997), a condition known as "sex-role reversal." Furthermore, the brooding structures used by males to gestate developing embryos represent an entirely novel and often complex reproductive tissue (Stolting and Wilson 2007). Indeed, the brooding structures in some lineages are highly vascularized (Carcupino et al. 1997), facilitate transfer of ions and nutrients from father to offspring (Ripley 2009; Ripley and Foran 2009), and may be involved in the mediation of postcopulatory sexual selection (Partridge et al. 2008; Braga Goncalves et al. 2010; Paczolt and Jones 2010; Mobley et al. 2011). Phylogenetic relationships for the focal taxa of this study are also well supported (Wilson et al. 2003), so an effective comparative genomic analysis of "male pregnancy" genes in multiple lineages, presumably different with respect to the influence of postcopulatory sexual selection, is possible.

In this study we examined the evolution of over 800 protein-coding DNA sequences expressed in pregnant male brooding tissue of four syngnathid species to

address two key issues. We first assessed whether genes expressed differentially in the male brooding structure during pregnancy evolve more rapidly at the amino acid level relative to genes whose expression patterns do not change with respect to pregnancy status. Based on transcriptome data from pregnant and non-pregnant brooding tissue from two pipefish species, genes expressed at higher levels during pregnancy demonstrated a higher nonsynonymous substitution rate (d*N*) than pregnancy-depressed or non-differentially expressed genes. Second, we used two reciprocally monophyletic species pairs to directly test the prediction that protein divergence should be faster in the polyandrous relative to monogamous lineage within both pairings. Our results suggest that the branch-specific ratio of nonsynonymous to synonymous substitution rates (d*N*/d*S*) is elevated for the polyandrous lineage in both comparisons, consistent with the sexual selection hypothesis. We also compared several evolutionary models for each gene by likelihood ratio test, identifying 10 male brood pouch genes that appear to have undergone adaptive divergence specifically in the two polyandrous lineages.

Materials and Methods

Sample preparation and Roche 454 GS FLX sequencing

We obtained wild-caught adult male Gulf pipefish (*Syngnathus scovelli*), dwarf seahorses (*Hippocampus zosterae*), wide-bodied pipefish (*Stigmatopora nigra*), and banded pipefish (*Corythoichthys intestinalis*) through our own collection efforts, those of our colleagues, and the aquarium fish trade. Animals were housed in 35–100-L volumes of seawater in biologically filtered tanks at 25 C for varying periods of time following

procurement, as was necessary to obtain both pregnant and non-pregnant individuals. All pregnant males used for the study were in early stages of pregnancy, within the first trimester of gestation, and were brooding embryos at or before “state 2” of Ripley and Foran’s pipefish developmental series (Ripley and Foran 2009). Each male was euthanized immediately prior to dissection with a lethal dose of MS222 buffered to physiological pH. All brooding structures, including “pouch,” “flap,” and ventrally suspended epithelial tissues were carefully and quickly excised from each animal and snap frozen at -80 C. Embryos were cautiously removed and discarded from pregnant males, and the remaining brooding tissue was rinsed with sterile water before freezing. Table 13 summarizes relevant sample information for each study species.

Table 13 Summary of Syngnathid Specimens Used to Generate 454 Data.

Species	Source	Number of individuals
Syngnathus scovelli	Texas, United States (authors)	5 pregnant, 5 non-pregnant
Hippocampus zosterae	Texas, United States (authors)	10 pregnant
Stigmatopora nigra	Victoria, Australia (K. Mobley & B. Wong)	5 pregnant
Corythoichthys intestinalis	Indonesia (aquarium trade)	1 pregnant, 1 non-pregnant

We homogenized the collected brooding tissues by pestle and isolated total RNA from each sample using TRIzol® Reagent (Invitrogen), in accordance with the

manufacturer's standard protocol. At this stage total RNA was pooled in equal amounts across multiple individuals in some cases (see Table 13) in order to obtain sufficient material (55 µg per library) for mRNA selection with the Oligotex® mRNA Mini Kit (Qiagen). We used 720 ng of the resulting mRNA from each library as template for cDNA synthesis with the SMART™ cDNA Library Construction Kit (Clontech). In general the manufacturer's reagents and LD PCR guidelines were followed, but a modified CDSIII/3' cDNA Synthesis Primer (5'- TAG AGG CCG AGG CGG CCG ACA TGT TTT GTT TTT TTT TCT TTT TTT TTT VN -3') and SuperScript® II Reverse Transcriptase (Invitrogen) were used in place of kit reagents. All steps following LD PCR were completed as described in the SMART™ protocol, without cloning. We sent 15 µg of cDNA from each of the six libraries (pregnant *S. scovelli*, non-pregnant *S. scovelli*, pregnant *C. intestinalis*, non-pregnant *C. intestinalis*, pregnant *H. zosterae*, and pregnant *S. nigra*) to the Michigan State University Research Technology Support Facility, where the libraries were multiplexed and sequenced in two runs on a Roche 454 GS FLX® sequencer using Titanium® chemistry. Non-pregnant libraries were not generated for *H. zosterae* due to insufficient RNA quantities, and no non-pregnant *S. nigra* samples were available.

De novo transcriptome assembly, alignment of orthologous coding sequences, and transcript abundance estimates

The 454 reads were trimmed to remove low quality regions, polyA tails, and cDNA synthesis artifacts using the highly customizable pipeline clean_reads, derived from the

ngs_backbone suite of bioinformatics tools (Blanca et al. 2011). After discarding all cleaned reads less than 50 nt, we performed four *de novo* transcriptome assemblies (one for each species) using the CLC Genomics Workbench® version 4.8 (CLC bio). After varying assembly parameter values to achieve optimal results, we set the k-mer size at 20, and selected the remapping option with similarity criterion set at 0.97, for all final assemblies.

Orthologous transcripts across the four species were identified using a “reciprocal best BLAST hit” criterion (Rivera et al. 1998), wherein the BLAST hit (Altschul et al. 1990) for each search with the highest bit-score is used to establish the “best hit.” This stringent approach stipulated that four-way orthologs were only obtained in the event that all 12 pairwise BLAST searches were reciprocally consistent. We aligned the sequences within each orthologous group to a “reference” protein-coding sequence, which was obtained from BLASTx (Camacho et al. 2009) queries of the NCBI non-redundant protein sequence database. We used the alignment software MACSE (Ranwez et al. 2011) to generate multiple sequence alignments. MACSE makes adjustments to preserve open reading frames in the face of rampant insertion and deletion errors encountered with 454 sequence data. We manually inspected all sequence alignments to ensure reasonable representations of the protein-coding sequence for each species.

We estimated per-contig transcript abundance in four of the libraries (pregnant *S. scovelli*, non-pregnant *S. scovelli*, pregnant *C. intestinalis*, non-pregnant *C. intestinalis*) by mapping sequencing reads from a given library back to its respective species’

assembly with the RNA-seq module in the CLC Genomics Workbench®. The number of uniquely mapped reads per contig was then used as a proxy of transcript abundance, and we used the R package DEGseq (Wang et al. 2010) to statistically compare transcript abundance between pregnant and non-pregnant libraries separately for the two species. For each comparison DEGseq conducts a likelihood ratio test to assess whether there is evidence for a difference between libraries in the proportion of reads mapped to a contig. We used the likelihood ratio test p-values from DEGseq to categorize each *S. scovelli* and *C. intestinalis* ortholog as "pregnancy-enriched," "pregnancy-depressed," or "non-differentially expressed," after setting a false discovery rate of 0.05 (Benjamini and Hochberg 1995).

Molecular evolutionary and statistical analyses

Branch-specific maximum likelihood estimates of d*N*, d*S*, and d*N*/d*S* were obtained from 806 protein-coding sequence alignments using the codeml "free-ratio" model within PAML 4.5 (Yang 1997). Figure 6 depicts phylogenetic relationships among the four focal species, along with species information relevant to the analyses presented here. All statistical tests were either performed in Microsoft Excel® (Microsoft Corp.) or JMP Pro® version 9 (SAS Institute Inc.). To test for different amino acid-changing substitution rates (d*N*) among genes falling into the three aforementioned expression categories we simply performed two Kruskal-Wallis Tests, one for the *S. scovelli* data and one for the *C. intestinalis* data. For these tests, branch-specific d*N* was the response variable of interest because it directly reflect protein divergence. We also performed two

side-by-side comparisons (one for the *S. scovelli – H. zosterae* clade, and one for the *S. nigra – C. intestinalis* clade) of polyandrous and monogamous branch-specific d*N*/d*S*

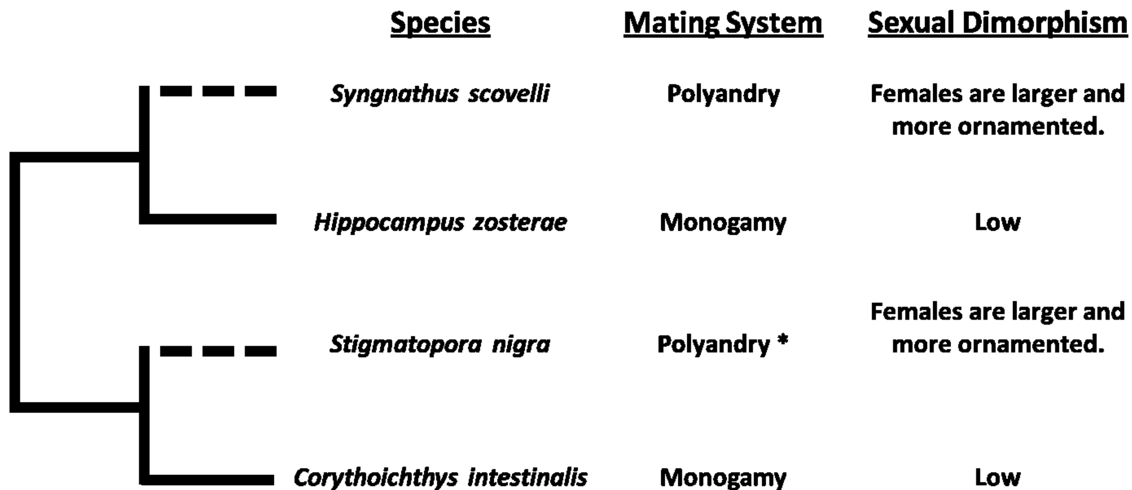

Figure 6 Mating System Variation Across Four Syngnathid Fish Species. Phylogenetic relationships among the four study species (adapted from Wilson et al. 2003), and key information regarding sexual selection (from Jones and Avise 2001, Masonjones and Lewis 2000, and Sogabe and Yanagisawa 2007). Dashed branches represent lineages along which sexual selection has intensified. *Polyandry is the suspected mating system for *S. nigra* based on unpublished observations.

ratios across all orthologs, using Wilcoxon signed-rank tests. In the context here, the d*N*/d*S* ratio is interpreted as a measure of amino acid divergence standardized by the nearly neutral "background" rate of silent substitution (Yang and Bielawski 2000). It is possible to obtain maximum likelihood estimates of d*N*/d*S* that are undefined when d*S* is equal to zero or effectively so. Due to the difficulty in interpreting these cases, we excluded from analysis all d*N*/d*S* estimates for genes lacking synonymous substitutions.

To assess the predicted role sexual selection might play in driving the divergence of proteins that are potentially relevant to male pregnancy we performed two routine likelihood ratio tests (LRTs) to compare codon models in PAML. First we performed the M7-M8 comparison for all alignments, which tests a model with a beta distribution of dN/dS values among codons in a sequence and no proportion of positively selected sites, against a model with the beta distribution and a class of codons for which $dN/dS > 1$ (Yang et al. 2000). If the positive selection model for a given alignment was significantly more likely than the null model according to the appropriate likelihood ratio test, we tentatively considered the gene to contain a class of sites with a history of positive selection. We then subjected this subset of "positively selected" genes to "branch-sites" tests for positive selection (Zhang et al. 2005), in which a class of positively selected sites along specified branches only distinguishes the alternative from the null model. We conducted the branch-sites test once, specifying polyandrous branches as foreground, and again specifying monogamous branches as foreground. In the case of a rejection of the null hypothesis in favor of the alternative, we recorded the number and identity of amino acids assigned to the positively selected class by the Bayes Empirical Bayes inference (Yang et al. 2005), requiring a minimum posterior probability of 0.95.

Results

454 sequencing, assembly, and orthology assignment

The two sequencing runs yielded a total of 1,772,265 passing reads with a mean read length of 281 nt. The number and mean length of reads for each library are reported in Table 14, along with species-specific assembly results. The four *de novo* assemblies resulted in a mean contig number of 13,254 and a mean contig length of 543 nt, on par with *de novo* transcriptome assemblies from similar studies (Elmer et al. 2010; Renaut et al. 2010). Our reciprocal best BLAST hit approach resulted in 848 orthologous groups across the four species. BLASTx queries of the NCBI nr database revealed that some of these groups consisted primarily of 5' or 3' UTRs, and several sequences lacked an open-reading frame altogether. After excluding these alignments from the data set, 806 orthologous groups with a mean alignment length of 126.4 codons remained for evolutionary analyses.

Table 14 Syngnathid 454 Sequencing and *de novo* Transcriptome Assemblies.
(P. = pregnant library; NP. = non-pregnant library).

Species	Read #	Read Length Mean (nt)	Contig #	Contig Length Mean (nt)
S. scovelli	P. 206,863	274	15,827	603
	NP. 373,152	281		
H. zosterae	P. 481,449	312	11,439	509
S. nigra	P. 194,818	301	8,868	540
C. intestinalis	P. 305,934	317	16,881	520
	NP. 210,049	200		

Pregnancy-related gene expression changes and dN in C. intestinalis *and* S. scovelli

We identified 122 putative "pregnancy-enriched," 125 "pregnancy-depressed," and 559 non-differentially expressed *C. intestinalis* orthologs. Likewise, we identified 39 putative "pregnancy-enriched," 80 "pregnancy-depressed," and 687 non-differentially expressed *S. scovelli* orthologs. For both species d*N* differs significantly among the three expression categories (*C. intestinalis* Kruskal-Wallis Test, $p = 0.0220$; *S. scovelli* Kruskal-Wallis Test, $p = 0.0003$). Pairwise comparisons *post hoc* revealed that pregnancy-enriched proteins have evolved more rapidly than pregnancy-depressed proteins in *C. intestinalis*, and pregnancy-enriched proteins have evolved more rapidly than non-differentially expressed proteins in *S. scovelli* (Bonferroni-corrected Mann-Whitney U tests). Comparisons of d*N*/d*S* yielded identical results, and there were no differences with respect to d*S* alone (results not shown). The mean, standard error of the mean, and median for each group are reported in Figure 7.

Two independent tests of mating system and branch-specific dN/dS

The lineage-specific maximum likelihood estimate of d*N*/d*S* across 806 genes expressed in pregnant male brooding tissue differs among the four species (Kruskal-Wallis Test, $p < 0.0001$, Figure 8); however, this test does not take advantage of the repeated d*N*/d*S* measures per gene, nor does it address whether mating system is in fact driving the species differences. We therefore performed two Wilcoxon signed-rank tests (one for each monogamous-polyandrous species pair), and discovered that in both cases the polyandrous-specific d*N*/d*S* is significantly higher on average than the monogamous-

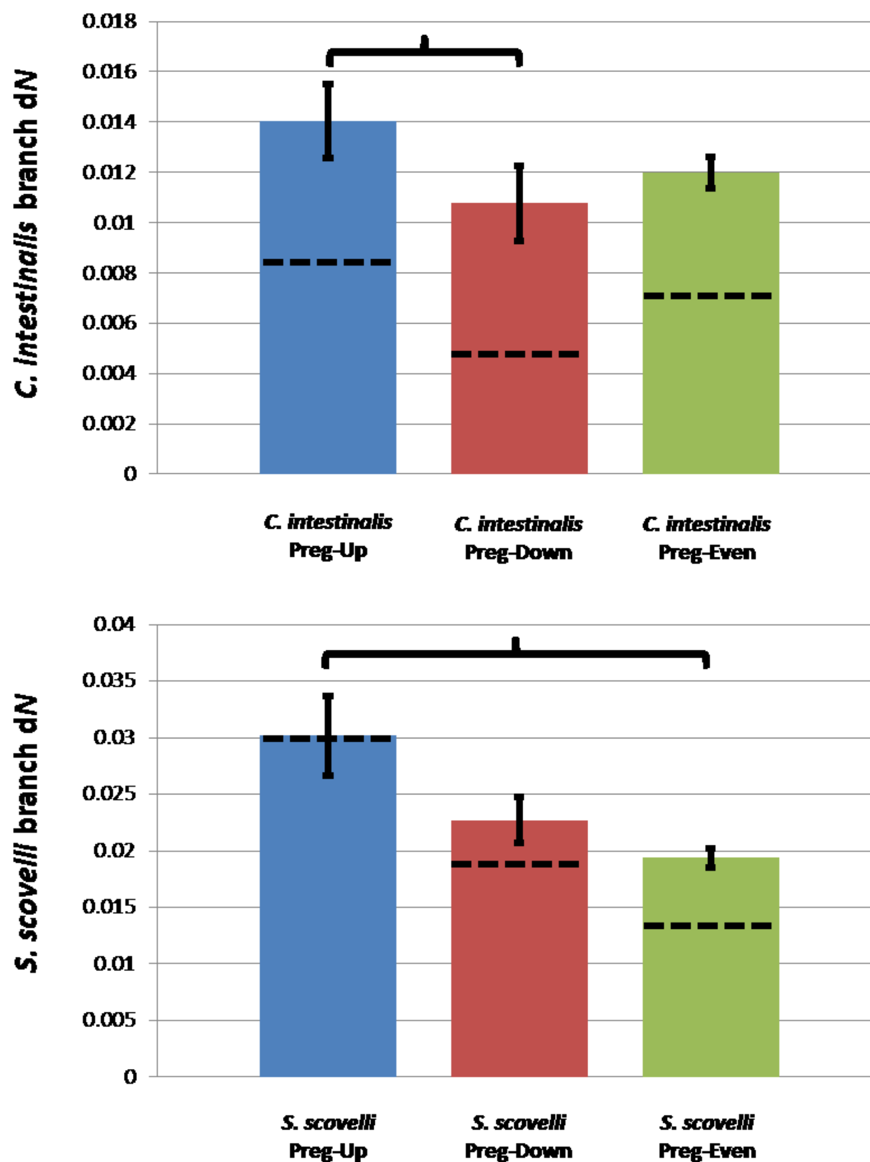

Figure 7 Expression Status and the Evolution of Male Brooding Proteins. The nonsynonymous substitution rate (d*N*) across pregnancy-enriched, pregnancy-depressed, and non-differentially expressed groups of genes in *C. intestinalis* (above) and *S. scovelli* (below). Colored bars represent the mean for each group, dashed lines mark the median, and error bars reflect standard errors of the mean. Any significant differences (Bonferroni-corrected $p < 0.05$) between group pairs are represented by brackets.

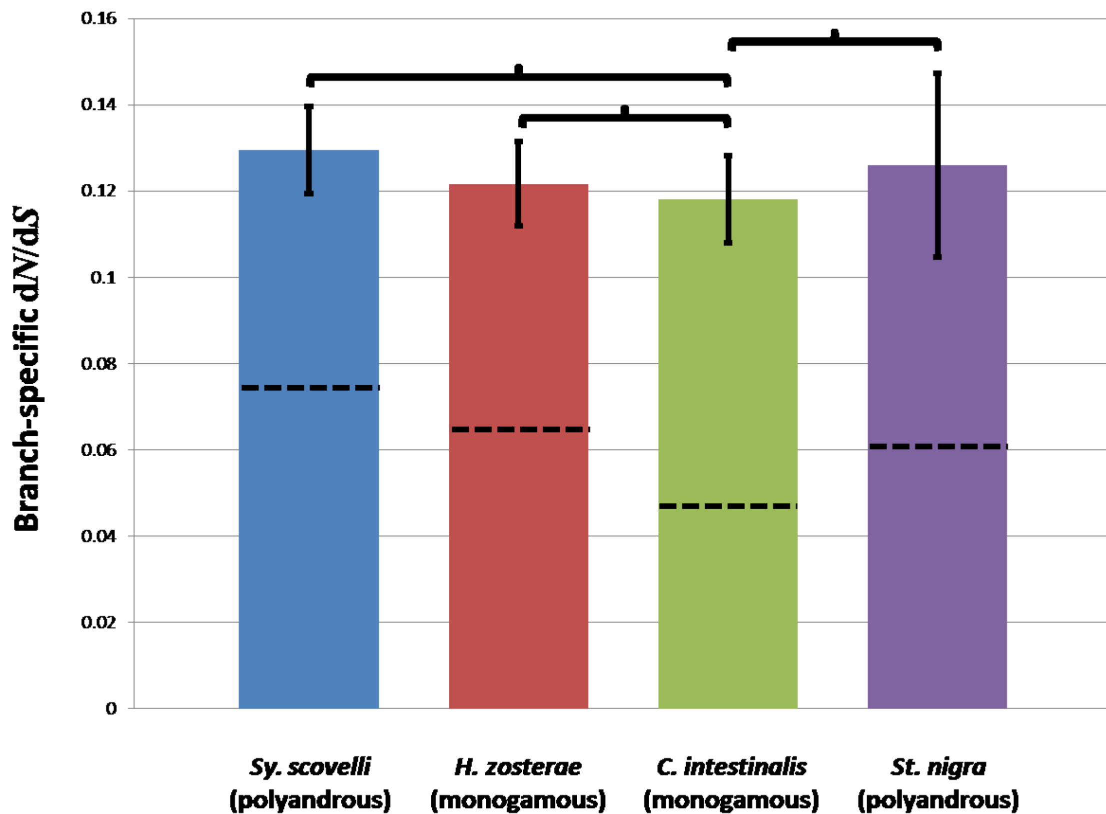

Figure 8 Branch-specific d*N*/d*S* Estimates Among Four Syngnathid Fish Species. The ratio of nonsynonymous to synonymous substitution rates (d*N*/d*S*) for each lineage in the study. Colored bars represent the mean d*N*/d*S* for each species across all 806 genes, dashed lines mark the median, and error bars reflect standard errors of the mean. Any significant differences (Bonferroni-corrected $p < 0.05$) between group pairs are represented by brackets.

specific d*N*/d*S* (*S. nigra*-*C. intestinalis* one-tailed test, $N = 753$ comparisons, $p = 0.0149$; *S. scovelli*-*H. zosterae* one-tailed test, $N = 772$ comparisons, $p = 0.0223$; Bonferroni-correction $\alpha/2 = 0.025$).

Sites and branch-sites likelihood ratio tests of positive selection

We initially carried out a M7-M8 likelihood ratio test for each of the 806 alignments to identify a subset of the genes for which there is some evidence of positive selection on a proportion of residues. A very small fraction (37 of 806) of the genes demonstrated a likelihood ratio test p-value less than 0.05. We did not set a false discovery rate for this panel of likelihood ratio tests because our intention was merely to identify liberally candidates for further analysis. This list of genes, including the identity of top BLASTx hits and relevant model comparison information, is reported in Table 15.

These 37 alignments were next subjected to 2 branch-sites likelihood ratio tests of positive selection. The first LRT yielded evidence for a class of positively selected sites along the two polyandrous branches in 10 of the 37 genes, and 11 individual amino acids were identified as belonging to the positively selected class (BEB p >0.95). The second LRT yielded evidence for a class of positively selected sites along the two monogamous branches in 8 of the 37 genes, and 6 individual amino acids were identified as belonging to the positively selected class (BEB $p > 0.95$). Table 15 presents these results and includes the identities of the residues belonging to the positively selected class for each model.

Table 15 Genes With Positively-selected Sites (M7-M8 Likelihood Ratio Test $p < 0.05$).
Residues under positive selection (BEB $p > 0.95$) in polyandrous (P) or monogamous (M) lineages are in bold.

Gene	BLASTx Hit Species	BLASTx Hit Acc. #	M7-M8 LRT *p*	Branch-Sites Positively-selected sites (P)	Branch-Sites Positively-selected sites (M)
14-3-3 zeta	*Artemia franciscana*	ABX80390	0.00001	**31 S, 168 T**	N.S. Branch-Sites LRT
60s ribosomal protein l4-a	*Pagrus major*	AAP20200	0.00002	**144 S, 199 G**	N.S. Branch-Sites LRT
keratin type II E3	*Epinephelus coioides*	AER42657	0.00003	**70 S**	**66 S**
type i cytoskeletal 13	*Anoplopoma fimbria*	ACQ58237	0.00005	**67 S**	**55 M, 83 Q**
c-type lectin 1	*Hippocampus comes*	AAQ56014	0.00024	N.S. Branch-Sites LRT	**None with p > 0.95**
cytochrome b-c1 complex subunit 7-like	*Tetraodon nigroviridis*	CAF99385	0.00054	**None with p > 0.95**	N.S. Branch-Sites LRT
protein s100-a1-like	*Oreochromis niloticus*	XP_003450931	0.00115	N.S. Branch-Sites LRT	N.S. Branch-Sites LRT
pdz and lim domain protein 2	*Salmo salar*	NP_001133275	0.00118	**145 N, 146 S**	N.S. Branch-Sites LRT
sh3 domain-containing ysc84-like protein 1-like	*Oreochromis niloticus*	XP_003458002	0.00211	N.S. Branch-Sites LRT	N.S. Branch-Sites LRT
sarcolemmal membrane-associated	*Oreochromis niloticus*	XP_003439061	0.00381	N.S. Branch-Sites LRT	N.S. Branch-Sites LRT
ribosomal protein s20	*Anoplopoma fimbria*	ACQ59028	0.00536	N.S. Branch-Sites LRT	**16 T**
ceramide synthase 5-like	*Oreochromis niloticus*	XP_003439062	0.00705	N.S. Branch-Sites LRT	N.S. Branch-Sites LRT
charged multivesicular body protein 2b-like	*Oreochromis niloticus*	XP_003445352	0.00797	N.S. Branch-Sites LRT	N.S. Branch-Sites LRT
peptidylprolyl isomerase like	*Oreochromis niloticus*	XP_003439534	0.00880	N.S. Branch-Sites LRT	N.S. Branch-Sites LRT
14-3-3 zeta	*Latrodectus hesperus*	ADV40156	0.01087	**None with p > 0.95**	N.S. Branch-Sites LRT
collagen type i alpha 3	*Oreochromis niloticus*	BAL40989	0.01101	**6 W, 38 S**	N.S. Branch-Sites LRT
tpa: nadph oxidase-1	*Oreochromis niloticus*	XP_003445197	0.01126	N.S. Branch-Sites LRT	**77 A**
aquaporin 3	*Dicentrarchus labrax*	ABG36519	0.01161	N.S. Branch-Sites LRT	N.S. Branch-Sites LRT

Table15 Continued

Gene	BLASTx Hit Species	BLASTx Hit Acc. #	M7-M8 LRT *p*	Branch-Sites Positively-selected sites (P)	Branch-Sites Positively-selected sites (M)
dna-directed rna polymerase i subunit rpa43	*Tetraodon nigroviridis*	CAG11135	0.01190	N.S. Branch-Sites LRT	N.S. Branch-Sites LRT
tetraspanin 8	*Osmerus mordax*	ACO09373	0.01360	**None with p > 0.95**	N.S. Branch-Sites LRT
rapunzel 2	*Danio rerio*	NP_001138713	0.01579	**89 P**	**None with p > 0.95**
type i keratin e7	*Oreochromis niloticus*	XP_003453824	0.01584	N.S. Branch-Sites LRT	**None with p > 0.95**
nucleolar protein 12	*Oreochromis niloticus*	XP_003443152	0.01646	N.S. Branch-Sites LRT	N.S. Branch-Sites LRT
barrier-to-autointegration factor	*Oreochromis niloticus*	XP_003450522	0.01682	N.S. Branch-Sites LRT	**48 E**
inositol monophosphatase 1-like	*Oreochromis niloticus*	XP_003439317	0.01687	N.S. Branch-Sites LRT	N.S. Branch-Sites LRT
ran-specific gtpase-activating protein	*Anoplopoma fimbria*	ACQ58002	0.01861	N.S. Branch-Sites LRT	N.S. Branch-Sites LRT
s100 calcium binding protein a10a	*Tetraodon nigroviridis*	CAG10829	0.01953	N.S. Branch-Sites LRT	N.S. Branch-Sites LRT
envoplakin-like	*Oreochromis niloticus*	XP_003443231	0.02088	N.S. Branch-Sites LRT	N.S. Branch-Sites LRT
14-3-3e1 protein	*Oreochromis niloticus*	XP_003456351	0.02221	N.S. Branch-Sites LRT	N.S. Branch-Sites LRT
myelin protein zero-like protein 2 precursor	*Oreochromis niloticus*	XP_003449346	0.02588	N.S. Branch-Sites LRT	N.S. Branch-Sites LRT
eukaryotic translation initiation factor 3 subunit e-b-like	*Oreochromis niloticus*	XP_003447252	0.02747	N.S. Branch-Sites LRT	N.S. Branch-Sites LRT
partitioning defective 6 homolog alpha-like	*Oreochromis niloticus*	XP_003449276	0.02968	N.S. Branch-Sites LRT	N.S. Branch-Sites LRT
costars family protein C6orf115	*Osmerus mordax*	ACO09013	0.03942	N.S. Branch-Sites LRT	N.S. Branch-Sites LRT
upf0552 protein c15orf38 homolog	*Oreochromis niloticus*	XP_003439446	0.03974	N.S. Branch-Sites LRT	N.S. Branch-Sites LRT
myosin regulatory light chain smooth muscle minor isoform-like	*Oreochromis niloticus*	XP_003439283	0.04571	N.S. Branch-Sites LRT	N.S. Branch-Sites LRT
ubiquitin carboxyl-terminal hydrolase 14-like	*Oreochromis niloticus*	XP_003438130	0.04828	N.S. Branch-Sites LRT	N.S. Branch-Sites LRT
cytolysin src-1-like	*Oreochromis niloticus*	XP_003442214	0.04892	N.S. Branch-Sites LRT	N.S. Branch-Sites LRT

Discussion

Variation in rate of protein divergence with respect to male pregnancy

Our results suggest that male brooding tissue proteins expressed at higher levels during pregnancy may evolve more rapidly than those under- or evenly-expressed during pregnancy relative to non-gestational periods (Figure 7). To some extent the effects of individual variation in transcript abundance unrelated to pregnancy were minimized in the *S. scovelli* comparison, because the pregnant and non-pregnant libraries were each derived from the pooled tissues of five males. The *C. intestinalis* comparison, however, is based on tissues from a single pregnant and a single non-pregnant male, which could explain the differences in expression group d*N* rank order between the two species, and the paucity of "differentially expressed" genes in *S. scovelli* relative to *C. intestinalis*.

Assuming the data accurately reflect gene expression differences between pregnant and non-pregnant libraries, an important consideration is whether d*N* is higher for differentially expressed genes because these genes play a functional role during pregnancy and have been subject to positive selection, or because, being more promiscuous in expression, they are also more likely to be under relaxed constraints. Genes expressed evenly across multiple tissues, for example, commonly demonstrate depressed nonsynonymous substitution rates (Duret and Mouchiroud 2000). A more thorough, replicated RNA-seq experiment comparing pregnant and non-pregnant males in several species would help discriminate between these two alternatives, because one would expect an equal degree of relaxed purifying selection on both pregnancy-enriched and pregnancy-depressed genes given the latter scenario. Tentatively, the data presented

here suggest that proteins up-regulated during early pregnancy belong to a rapidly-evolving class of male brooding tissue genes. If major pregnancy-induced cellular changes in the male brooding structure are closely tied to offspring fitness (Carcupino et al. 1997; Ripley and Foran 2009), the genes associated with these changes have likely experienced natural and/or sexual selection in the past or present.

Elevated dN/dS in polyandrous syngnathid lineages

Previous studies assessing the predicted relationship between mating system and rates of reproductive molecular evolution have generally suffered at least one of several limitations. A substantial barrier, which inherently precludes the understanding of a phenomenon that is predicted to affect at least hundreds of loci per genome, is simply the small number of genes analyzed. In a recent review summarizing the evidence for a relationship between mating system and rates of reproductive gene evolution, Wong (Wong 2011) reported that the largest number of genes examined in a given study of this nature is 13 (Finn and Civetta 2010). Another barrier is a lack of sufficient mating system variation, or limited information on the precise nature of the history of sexual selection in focal taxa. Indeed, an important requirement for the ideal framework in which to test the sexual selection hypothesis is the existence of multiple, independent changes to a derived mating system along the phylogeny, exemplified by a study of 20 rodent species (Ramm et al. 2008).

Though our study includes only four species, we analyzed over 800 protein-coding sequences, and our choice of taxa allowed for two independent comparisons of

polyandrous versus monogamous mating systems. The presumed ancestral mating system for syngnathid fishes, as inferred through parsimony, is monogamy (Wilson et al. 2003). Since divergence of the four lineages examined in our study, ancestors of *Syngnathus* and *Stigmatopora* species underwent independent transitions from monogamous to polyandrous mating systems, while *Hippocampus* and *Corythoichthys* presumably retained ancestral monogamy (Figure 6). As noted, this arrangement allows for two separate tests of the prediction that lineages with a history of more intense postcopulatory sexual selection should demonstrate more rapid rates of reproductive protein evolution. Indeed, we found that this prediction holds for both of our comparisons. *Syngnathus*-specific d*N*/d*S* is consistently higher than *Hippocampus*-specific d*N*/d*S*, and *Stigmatopora*-specific d*N*/d*S* is similarly higher than *Corythoichthys*-specific d*N*/d*S* across several hundred genes expressed in the male brooding tissue during pregnancy. It does appear that the *Syngnathus*- *Hippocampus* dichotomy may be weaker than the *Stigmatopora*- *Corythoichthys* difference (evident in Figure 8), which could be due to highly derived brood pouch complexity in seahorse ancestors (Carcupino et al. 2002; Stolting and Wilson 2007). *Corythoichthys* pipefishes, on the other hand, have the least derived brooding morphology of the four genera in our study (Dawson 1985), so future investigation should address the relative contributions of mating system and brooding structure evolution to the divergence of male pregnancy proteins.

A higher gene-wide d*N*/d*S* ratio in a given comparison among lineages does not necessarily indicate protein modification through positive natural selection, because

lineage-specific relaxation of purifying selection can produce the same pattern (Yang 1998; Fay and Wu 2003). In light of this, we must be somewhat cautious when interpreting the true selective causes of observed lineage-specific patterns. In our study the fact that the predicted pattern was consistent for two polyandrous-monogamous comparisons is reassuring, but two historical relaxations of constraint are still within the realm of possibility. This explanation becomes especially pertinent when one considers that highly skewed mating systems are associated with lower effective population sizes (Crow and Kimura 1970). Reductions in effective population size along a lineage can reduce the efficacy with which natural selection purges slightly deleterious mutations (Ohta 1973), which may result in a higher fixation rate of mildly deleterious nonsynonymous substitutions in lineages with highly skewed mating systems. This phenomenon should affect the entire genome, however, so future comparisons between mating system- d*N*/d*S* associations for reproductive and non-reproductive proteins should help discriminate between alternatives.

We performed likelihood ratio tests to compare positive selection codon models with null alternatives for each gene. Our initial comparison of M7 vs. M8, a relatively liberal likelihood ratio test, revealed some evidence for a proportion of positively selected amino acids in 37 of the alignments (Table 15). This is a small fraction of the 806 brooding structure genes we analyzed, but the low number could reflect low power of sites tests conducted on alignments of just four sequences. We fully expect analyses in the near future to include data from a dozen or so syngnathid species, which should greatly increase the power of these methods to detect sites under positive selection.

Branch-sites tests (Zhang et al. 2005) for these 37 candidates were carried out to assess the evidence for positive selection first on a class of sites in both polyandrous lineages, and again in both monogamous lineages. As Table 15 indicates, there appears to be no glaring disparity in the frequency of genes experiencing positive selection, or in the number of positively selected residues, when considering polyandrous versus monogamous lineages. Again, this result could be in part due to the evolution of increased pouch complexity in the lineage leading to *Hippocampus*, but further study involving more taxa, variable in both mating system and brooding morphology, is clearly warranted.

In summary, our results indicate that male brooding tissue proteins expressed differentially during pregnancy evolve more rapidly than those expressed statically. Surprisingly, and in accordance with the sexual selection hypothesis for the rapid evolution of reproductive proteins, male brooding tissue genes diverge more rapidly in polyandrous relative to monogamous lineages of syngnathid fishes. It should be noted that other modes of molecular evolution, particularly transcriptome-wide turnover in genes recruited for reproductive function, may be an equally or more important consequence of postcopulatory sexual selection. Some, for example, have established evidence for rapid evolution of seminal fluid proteome composition among muroid rodents (Ramm et al. 2009). It has already been established that genes such as astacin metalloproteases, historically transcribed in kidney and liver, have been coopted for expression in the brood pouch of *Syngnathus* pipefishes (Harlin-Cognato et al. 2006). There may be substantial differences among lineages in the potpourri of molecules

recruited for expression in this novel reproductive tissue, and the extent to which postcopulatory sexual selection might drive such differences is an important and interesting question. Answers to this question, and more complete data with respect to the coding sequence evolution of male pregnancy genes, are well within grasp thanks to the ever-transforming status of high-throughput DNA sequencing technologies. Both thorough reappraisals of the preliminary results described here, and new insights into the evolutionary genomics of reproduction in syngnathid fishes and beyond, are on the immediate horizon.

5. SUMMARY

To say the present is an exciting time for genome sciences is an understatement. As is the case for many questions in molecular biology and evolution, genomic insights into sexual reproduction are unfolding at an unprecedented rate thanks to the incipient transformation of DNA sequencing technologies. One needs to look no further than the contents of this dissertation to understand the remarkable transition in methodologies that has recently taken biology by storm. In Section 2 of this document the issue of sex-biased gene expression was addressed using microarray technology and a model vertebrate, the zebrafish, for which this particular tool was commercially tailored. Sections 3 and 4 describe the tackling of very similar questions, but they feature massively parallel (Roche 454®) cDNA sequencing as the means to gain insight, a tool that permits the simultaneous acquisition of sequence and expression data along with a relaxed constraint on the choice of study organism. It is now even apparent that, except for special applications, 454 sequencing is being replaced as the technology of choice by much higher throughput sequencing platforms.

I make this point because it is important to understand a few things about the nature of the data and results published in this dissertation. Data quality issues are at hand whenever new approaches are applied, and a lack of experience with results from these approaches means interpretations of the data are made without a full realization of the technological shortcomings. Another quandary associated with transitional science is that imminent techniques far superior to the current ones will most likely adjust if not

altogether shatter many present conclusions in the very near future. Keeping these issues in mind, the conclusions presented in the preceding pages should be viewed as tentative and subject to change in the face of new and higher-volume data.

Disclaimers aside, the contents of the three main sections within this document do contribute to the burgeoning knowledge of how sex-specific selective processes affect animal genomes. The primary aims of the dissertation were 1.) to review the current literature regarding sexual selection and the evolution of reproductive molecules (Section 1), 2.) to appraise the nature of sex-biased gene expression in organisms other than those species already examined in this regard (Sections 2 and 3), 3.) to characterize the relationship between spatial expression patterns of genes and their rates of sequence evolution (Sections 3 and 4), and 4.) to explore the possibility that postcopulatory sexual selection drives the rapid evolution of male pregnancy genes, a novel class of reproductive molecules (Section 4). Teleost fishes were used as study subjects for a variety of reasons, including convenience and available resources, but most importantly because of unique taxon-specific attributes with respect to reproductive biology. Up until this point our understanding of reproductive protein evolution and sexual selection at molecular levels was based almost entirely on data from *Drosophila*, rodents, great apes, and a few miscellaneous birds, other arthropods, and marine invertebrates. Through the exploitation of the teleosts, a group extremely diverse in species, life history, morphology, and behavior, one is able to greatly extend generalizations drawn from traditional molecular biological model systems.

To summarize the results from Section 2, male and female zebrafish demonstrate a great deal of sex-biased gene expression (nearly 40% of the genome), a substantial proportion considering that major genomic differences between male and female zebrafish do not exist (Bradley et al. 2011). All statistically well-supported sex differences were based on the ovary-testis comparison, and we found almost no sex-biases in expression comparing gonad-dissected bodies. Furthermore, we discovered more male-biased than female-biased genes, and male-biased genes as a group were differentially expressed at higher levels than were female-biased genes.

In Section 3 I described a next-generation sequencing study of sex-, tissue-, and species-biased gene expression in two naturally hybridizing swordtail species (genus *Xiphophorus*). As expected, much of the transcriptome for each tissue analyzed (male and female gonad, head, and body) was found to be sex-biased in both species. Somewhat at odds with respect to sex-biased gene expression patterns in other vertebrates, we found an excess of female-biased expression for gonads, and an excess of male-biased expression for head and body tissues. Also, sex-biased genes in *Xiphophorus* diverge more rapidly than non-sex-biased genes, particularly when the focus is on reproductive tissues. Regarding tissue-specific gene expression and sequence divergence, we found that tissue specificity likely has some bearing on whether or not coding sequences of genes evolve rapidly. We found, however, that rates of sequence divergence for gonad-biased genes are no higher than rates of divergence for head-biased genes, an important result. Finally, an assessment of the relationship between species differences in gene expression and coding sequence evolution revealed that

genes more divergent in expression between *X. birchmanni* and *X. malinche* also tend to be more differentiated at the level of the coding sequence, but only when species-biased genes in the testes and female head are considered.

A novel reproductive tissue, the male brooding structure of syngnathid fishes, was the focus of the study described in Section 4. High-throughput pyrosequencing of cDNA libraries derived from brooding tissues of pregnant and non-pregnant males in two pipefish species (*Corythoichthys intestinalis* and *Syngnathus scovelli*) revealed that genes up- and possibly down-regulated during pregnancy evolve rapidly at the amino acid level. Perhaps most interesting, however, is the evidence that branch-specific rates of functional coding sequence evolution are higher for lineages presumed to have experienced strong sexual selection in the past and present, relative to monogamous lineages. This result, based on roughly 800 protein-coding genes expressed in the pregnant male brood pouch, is consistent with the elusive hypothesis that postcopulatory sexual selection drives the rapid evolution of reproductive molecules.

Much work remains to be done in these three teleost groups if they are to contribute substantially to our knowledge of reproductive molecular evolution. An especially useful expansion of the studies presented in this dissertation would simply involve the addition of high quality sequencing data from many more species. The power to detect real signals of selection in sequence data increases greatly with the number of species sampled. Another improvement would be next-generation expression data from multiple individuals per treatment or species of interest, so that variation in transcript abundance that truly reflects biological processes such as pregnancy status can

be captured. Fortunately these improvements are well within reach, thanks to the ever-increasing power, and ever-decreasing cost, of DNA sequencing methods. It is my hope that the initial observations reported here and elsewhere contribute to a “first wave” of change in how we approach key questions in evolutionary genomics, and that the rich data yet to come truly transform our understanding of evolution in molecular terms.

REFERENCES

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

Andersson, MB. 1994. Sexual Selection: Princeton University Press, Princeton.

Andres, JA, LS Maroja, SM Bogdanowicz, WJ Swanson, RG Harrison. 2006. Molecular evolution of seminal proteins in field crickets. *Mol Biol Evol* 23:1574-1584.

Arbeitman, MN, EE Furlong, F Imam, E Johnson, BH Null, BS Baker, MA Krasnow, MP Scott, RW Davis, KP White. 2002. Gene expression during the life cycle of Drosophila melanogaster. *Science* 297:2270-2275.

Arnegard, ME, PB McIntyre, LJ Harmon, ML Zelditch, WG Crampton, JK Davis, JP Sullivan, S Lavoue, CD Hopkins. 2010. Sexual signal evolution outpaces ecological divergence during electric fish species radiation. *American Naturalist* 176:335-356.

Arnqvist, G. 1998. Comparative evidence for the evolution of genitalia by sexual selection. *Nature* 393:784-786.

Artieri, CG, W Haerty, BP Gupta, RS Singh. 2008. Sexual selection and maintenance of sex: evidence from comparisons of rates of genomic accumulation of mutations and divergence of sex-related genes in sexual and hermaphroditic species of Caenorhabditis. *Mol Biol Evol* 25:972-979.

Baldi, P, AD Long. 2001. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* 17:509-519.

Bardin, CW, JF Catterall. 1981. Testosterone - a major determinant of extra-genital sexual dimorphism. *Science* 211:1285-1294.

Bateman, AJ. 1948. Intra-sexual selection in Drosophila. *Heredity (Edinb)* 2:349-368.

Benjamini, Y, Y Hochberg. 1995. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Methodological* 57:289-300.

Blanca, JM, L Pascual, P Ziarsolo, F Nuez, J Canizares. 2011. ngs_backbone: a pipeline for read cleaning, mapping and SNP calling using next generation sequence. *BMC Genomics* 12.

Boschetto, C, C Gasparini, A Pilastro. 2011. Sperm number and velocity affect sperm competition success in the guppy (Poecilia reticulata). *Behavioral Ecology and Sociobiology* 65:813-821.

Bradley, KM, JP Breyer, DB Melville, KW Broman, EW Knapik, JR Smith. 2011. An SNP-based linkage map for zebrafish reveals sex determination loci. *G3 (Bethesda)* 1:3-9.

Braga Goncalves, I, KB Mobley, I Ahnesjo, G Sagebakken, AG Jones, C Kvarnemo. 2010. Reproductive compensation in broad-nosed pipefish females. *Proceedings of the Royal Society B-Biological Sciences* 277:1581-1587.

Breedlove, SM. 1992. Sexual dimorphism in the vertebrate nervous system. *Journal of Neuroscience* 12:4133-4142.

Calkins, JD, D El-Hinn, WJ Swanson. 2007. Adaptive evolution in an avian reproductive protein: ZP3. *J Mol Evol* 65:555-563.

Camacho, C, G Coulouris, V Avagyan, N Ma, J Papadopoulos, K Bealer, TL Madden. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.

Carcupino, M, A Baldacci, M Mazzini, P Franzoi. 1997. Morphological organization of the male brood pouch epithelium of Syngnathus abaster Risso (Teleostea, Syngnathidae) before, during, and after egg incubation. *Tissue & Cell* 29:21-30.

Carcupino, M, A Baldacci, M Mazzini, P Franzoi. 2002. Functional significance of the male brood pouch in the reproductive strategies of pipefishes and seahorses: a morphological and ultrastructural comparative study on three anatomically different pouches. *Journal of Fish Biology* 61:1465-1480.

Carney, GE. 2007. A rapid genome-wide response to Drosophila melanogaster social interactions. *BMC Genomics* 8:288.

Charlesworth, B. 1991. The evolution of sex-chromosomes. *Science* 251:1030-1033.

Chuong, EB, W Tong, HE Hoekstra. 2010. Maternal-fetal conflict: rapidly evolving proteins in the rodent placenta. *Mol Biol Evol* 27:1221-1225.

Civetta, A, RS Singh. 1995. High divergence of reproductive tract proteins and their association with postzygotic reproductive isolation in Drosophila melanogaster and Drosophila virilis group species. *J Mol Evol* 41:1085-1095.

Connallon, T, LL Knowles. 2005. Intergenomic conflict revealed by patterns of sex-biased gene expression. *Trends in Genetics* 21:495-499.

Coulthart, MB, RS Singh. 1988. High level of divergence of male-reproductive-tract proteins, between Drosophila melanogaster and its sibling species, D. simulans. *Mol Biol Evol* 5:182-191.

Crow, JF, M Kimura. 1970. An Introduction to Population Genetics Theory: Blackburn Press, Caldwell.

Cutter, AD, S Ward. 2005. Sexual and temporal dynamics of molecular evolution in C. elegans development. *Mol Biol Evol* 22:178-188.

Darwin, C. 1859. On the Origin of the Species by Means of Natural Selection: Or, the Preservation of Favoured Races in the Struggle for Life. 2009 edition: Cambridge University Press, Cambridge.

Darwin, C. 1871. The Descent of Man, and Selection in Relation to Sex. 2009 edition: Cambridge University Press, Cambridge.

Dawson, CE. 1985. Indo-Pacific Pipefishes: Red Sea to the Americas: Gulf Coast Research Laboratory, Ocean Springs.

de Carvalho, CE, H Tanaka, N Iguchi, S Ventela, H Nojima, Y Nishimune. 2002. Molecular cloning and characterization of a complementary DNA encoding sperm tail protein SHIPPO 1. *Biol Reprod* 66:785-795.

Del Giacco, L, S Diani, F Cotelli. 2000. Identification and spatial distribution of the mRNA encoding an egg envelope component of the cyprinid zebrafish, Danio rerio, homologous to the mammalian ZP3(ZPC). *Dev Genes Evol* 210:41-46.

Duret, L, D Mouchiroud. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol Biol Evol* 17:68-74.

Eberhard, WG. 2009. Postcopulatory sexual selection: Darwin's omission and its consequences. *Proc Natl Acad Sci U S A* 106:10025-10032.

Eberhard, WG. 2011. Experiments with genitalia: a commentary. *Trends Ecol Evol* 26:17-21.

Ellegren, H, L Hultin-Rosenberg, B Brunstrom, L Dencker, K Kultima, B Scholz. 2007. Faced with inequality: chicken do not have a general dosage compensation of sex-linked genes. *BMC Biol* 5:40.

Ellegren, H, J Parsch. 2007. The evolution of sex-biased genes and sex-biased gene expression. *Nat Rev Genet* 8:689-698.

Elmer, KR, S Fan, HM Gunter, JC Jones, S Boekhoff, S Kuraku, A Meyer. 2010. Rapid evolution and selection inferred from the transcriptomes of sympatric crater lake cichlid fishes. *Mol Ecol* 19 Suppl 1:197-211.

Ewert, MA, DR Jackson, CE Nelson. 1994. Patterns of temperature-dependent sex determination in turtles. *Journal of Experimental Zoology* 270:3-15.

Falgueras, J, AJ Lara, N Fernandez-Pozo, FR Canton, G Perez-Trabado, MG Claros. 2010. SeqTrim: a high-throughput pipeline for pre-processing any type of sequence read. *BMC Bioinformatics* 11:38.

Fay, JC, CI Wu. 2003. Sequence divergence, functional constraint, and selection in protein evolution. *Annu Rev Genomics Hum Genet* 4:213-235.

Finn, S, A Civetta. 2010. Sexual selection and the molecular evolution of ADAM Proteins. *J Mol Evol* 71:231-240.

Fisher, HS, GG Rosenthal. 2006. Female swordtail fish use chemical cues to select well-fed mates. *Anim Behav* 72:721-725.

Fisher, HS, GG Rosenthal. 2007. Male swordtails court with an audience in mind. *Biol Lett* 3:5-7.

Geyer, LB, SR Palumbi. 2003. Reproductive character displacement and the genetics of gamete recognition in tropical sea urchins. *Evolution* 57:1049-1060.

Good, JM, MW Nachman. 2005. Rates of protein evolution are positively correlated with developmental timing of expression during mouse spermatogenesis. *Mol Biol Evol* 22:1044-1052.

Grath, S, J Parsch. 2012. Rate of amino acid substitution is influenced by the degree and conservation of male-biased transcription over 50 million years of Drosophila evolution. *Genome Biol Evol*.

Haerty, W, S Jagadeeshan, RJ Kulathinal, et al. 2007. Evolution in the fast lane: rapidly evolving sex-related genes in Drosophila. *Genetics* 177:1321-1335.

Hahn, MW, GC Lanzaro. 2005. Female-biased gene expression in the malaria mosquito Anopheles gambiae. *Curr Biol* 15:R192-193.

Hamilton, WD, M Zuk. 1989. Parasites and sexual selection - Reply. *Nature* 341:289-290.

Harlin-Cognato, A, EA Hoffman, AG Jones. 2006. Gene cooption without duplication during the evolution of a male-pregnancy gene in pipefish. *Proc Natl Acad Sci U S A* 103:19407-19412.

Herlyn, H, H Zischler. 2007. Sequence evolution of the sperm ligand zonadhesin correlates negatively with body weight dimorphism in primates. *Evolution* 61:289-298.

Hosken, DJ, P Stockley. 2004. Sexual selection and genital evolution. *Trends Ecol Evol* 19:87-93.

Hsiao, CD, HJ Tsai. 2003. Transgenic zebrafish with fluorescent germ cell: a useful tool to visualize germ cell proliferation and juvenile hermaphroditism in vivo. *Dev Biol* 262:313-323.

Hubbell, E, WM Liu, R Mei. 2002. Robust estimators for expression analysis. *Bioinformatics* 18:1585-1592.

Hurle, B, W Swanson, ED Green. 2007. Comparative sequence analyses reveal rapid and divergent evolutionary changes of the WFDC locus in the primate lineage. *Genome Res* 17:276-286.

Jiang, M, J Ryu, M Kiraly, K Duke, V Reinke, SK Kim. 2001. Genome-wide analysis of developmental and sex-regulated gene expression profiles in Caenorhabditis elegans. *Proc Natl Acad Sci U S A* 98:218-223.

Jin, W, RM Riley, RD Wolfinger, KP White, G Passador-Gurgel, G Gibson. 2001. The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster. *Nat Genet* 29:389-395.

Jones, AG, JC Avise. 1997. Microsatellite analysis of maternity and the mating system in the Gulf pipefish Syngnathus scovelli, a species with male pregnancy and sex-role reversal. *Mol Ecol* 6:203-213.

Jones, AG, JC Avise. 2001. Mating systems and sexual selection in male-pregnant pipefishes and seahorses: insights from microsatellite-based studies of maternity. *Journal of Heredity* 92:150-158.

Jordan, IK, L Mariño-Ramírez, YI Wolf, EV Koonin. 2004. Conservation and coevolution in the scale-free human gene coexpression network. *Mol Biol Evol* 21:2058-2070.

Khaitovich, P, I Hellmann, W Enard, K Nowick, M Leinweber, H Franz, G Weiss, M Lachmann, S Pääbo. 2005. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science* 309:1850-1854.

Lande, R. 1980. Sexual dimorphism, sexual selection, and adaptation in polygenic characters. *Evolution* 34:292-305.

Lemos, B, BR Bettencourt, CD Meiklejohn, DL Hartl. 2005. Evolution of proteins and gene expression levels are coupled in Drosophila and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. *Mol Biol Evol* 22:1345-1354.

Li, C, W Hung Wong. 2001. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol* 2:RESEARCH0032.

Li, C, WH Wong. 2001. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci U S A* 98:31-36.

Li, Y, JM Chia, R Bartfai, A Christoffels, GH Yue, K Ding, MY Ho, JA Hill, E Stupka, L Orban. 2004. Comparative analysis of the testis and ovary transcriptomes in zebrafish by combining experimental and computational tools. *Comp Funct Genomics* 5:403-418.

Liu, WM, R Mei, X Di, et al. 2002. Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics* 18:1593-1599.

Liu, XJ, H Wang, ZY Gong. 2006. Tandem-repeated zebrafish zp3 genes possess oocyte-specific promoters and are insensitive to estrogen induction. *Biol Reprod* 74:1016-1025.

Lu, J, C-I Wu. 2005. Weak selection revealed by the whole-genome comparison of the X chromosome and autosomes of human and chimpanzee. *Proc Natl Acad Sci U S A* 102:4063-4067.

Malone, JH, DL Hawkins, Jr., P Michalak. 2006. Sex-biased gene expression in a ZW sex determination system. *J Mol Evol* 63:427-436.

Mank, JE, E Axelsson, H Ellegren. 2007a. Fast-X on the Z: rapid evolution of sex-linked genes in birds. *Genome Res* 17:618-624.

Mank, JE, L Hultin-Rosenberg, E Axelsson, H Ellegren. 2007b. Rapid evolution of female-biased, but not male-biased, genes expressed in the avian brain. *Mol Biol Evol* 24:2698-2706.

Mank, JE, L Hultin-Rosenberg, MT Webster, H Ellegren. 2008. The unique genomic properties of sex-biased genes: insights from avian microarray data. *BMC Genomics* 9:148.

Martin-Coello, J, H Dopazo, L Arbiza, J Ausio, ER Roldan, M Gomendio. 2009. Sexual selection drives weak positive selection in protamine genes and high promoter divergence, enhancing sperm competitiveness. *Proc Biol Sci* 276:2427-2436.

Masonjones, HD, SM Lewis. 2000. Differences in potential reproductive rates of male and female seahorses related to courtship roles. *Anim Behav* 59:11-20.

Masta, SE, WP Maddison. 2002. Sexual selection driving diversification in jumping spiders. *Proc Natl Acad Sci U S A* 99:4442-4447.

Meiklejohn, CD, J Parsch, JM Ranz, DL Hartl. 2003. Rapid evolution of male-biased gene expression in Drosophila. *Proc Natl Acad Sci U S A* 100:9894-9899.

Meisel, RP. 2011. Towards a more nuanced understanding of the relationship between sex-biased gene expression and rates of protein-coding sequence evolution. *Mol Biol Evol* 28:1893-1900.

Metz, EC, SR Palumbi. 1996. Positive selection and sequence rearrangements generate extensive polymorphism in the gamete recognition protein bindin. *Mol Biol Evol* 13:397-406.

Mobley, KB, C Kvarnemo, I Ahnesjo, C Partridge, A Berglund, AG Jones. 2011. The effect of maternal body size on embryo survivorship in the broods of pregnant male pipefish. *Behavioral Ecology and Sociobiology* 65:1169-1177.

Mortazavi, A, BA Williams, K Mccue, L Schaeffer, B Wold. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* 5:621-628.

Nuzhdin, SV, ML Wayne, KL Harmon, LM McIntyre. 2004. Common pattern of evolution of gene expression level and protein sequence in Drosophila. *Mol Biol Evol* 21:1308-1317.

Ohta, T. 1973. Slightly deleterious mutant substitutions in evolution. *Nature* 246:96-98.

Paczolt, KA, AG Jones. 2010. Post-copulatory sexual selection and sexual conflict in the evolution of male pregnancy. *Nature* 464:401-U494.

Palmer, CA, RA Watts, LD Houck, AL Picard, SJ Arnold. 2007. Evolutionary replacement of components in a salamander pheromone signaling complex: more evidence for phenotypic-molecular decoupling. *Evolution* 61:202-215.

Parisi, M, R Nuttall, P Edwards, et al. 2004. A survey of ovary-, testis-, and soma-biased gene expression in Drosophila melanogaster adults. *Genome Biol* 5.

Parisi, M, R Nuttall, D Naiman, G Bouffard, J Malley, J Andrews, S Eastman, B Oliver. 2003. Paucity of genes on the Drosophila X chromosome showing male-biased expression. *Science* 299:697-700.

Parker, GA. 1970. Sperm competition and its evolutionary consequences in insects. *Biological Reviews of the Cambridge Philosophical Society* 45:525-&.

Parker, GA, L Partridge. 1998. Sexual conflict and speciation. *Philosophical Transactions of the Royal Society B-Biological Sciences* 353:261-274.

Partridge, C, I Ahnesjö, C Kvarnemo, KB Mobley, A Berglund, AG Jones. 2008. The effect of perceived female parasite load on post-copulatory male choice in a sex-role-reversed pipefish. *Behavioral Ecology and Sociobiology* 63:345-354.

Pilastro, A, JP Evans, S Sartorelli, A Bisazza. 2002. Male phenotype predicts insemination success in guppies. *Proc Biol Sci* 269:1325-1330.

Pilastro, A, M Simonato, A Bisazza, JP Evans. 2004. Cryptic female preference for colorful males in guppies. *Evolution* 58:665-669.

Podlaha, O, J Zhang. 2003. Positive selection on protein-length in the evolution of a primate sperm ion channel. *Proc Natl Acad Sci U S A* 100:12241-12246.

Prince, EG, D Kirkland, JP Demuth. 2010. Hyperexpression of the X chromosome in both sexes results in extensive female bias of X-linked genes in the flour beetle. *Genome Biol Evol* 2:336-346.

Rajagopalan, D. 2003. A comparison of statistical methods for analysis of high density oligonucleotide array data. *Bioinformatics* 19:1469-1476.

Ramm, SA, L McDonald, JL Hurst, RJ Beynon, P Stockley. 2009. Comparative proteomics reveals evidence for evolutionary diversification of rodent seminal fluid and its functional significance in sperm competition. *Mol Biol Evol* 26:189-198.

Ramm, SA, PL Oliver, CP Ponting, P Stockley, RD Emes. 2008. Sexual selection and the adaptive evolution of mammalian ejaculate proteins. *Mol Biol Evol* 25:207-219.

Ranwez, V, S Harispe, F Delsuc, EJP Douzery. 2011. MACSE: Multiple alignment of coding sequences accounting for frameshifts and stop codons. *PLoS One* 6:e22594.

Ranz, JM, CI Castillo-Davis, CD Meiklejohn, DL Hartl. 2003. Sex-dependent gene expression and evolution of the Drosophila transcriptome. *Science* 300:1742-1745.

Rauchenberger, M, KD Kallman, DC Morizot. 1990. Monophyly and geography of the Rio Pa´nuco basin swordtails (genus Xiphophorus) with descriptions of four new species. *American Museum Novitates* 2975:1-41.

Renaut, S, AW Nolte, L Bernatchez. 2010. Mining transcriptome sequences towards identifying adaptive single nucleotide polymorphisms in lake whitefish species pairs (Coregonus spp. Salmonidae). *Mol Ecol* 19 Suppl 1:115-131.

Rinn, JL, JS Rozowsky, IJ Laurenzi, PH Petersen, K Zou, W Zhong, M Gerstein, M Snyder. 2004. Major molecular differences between mammalian sexes are involved in drug metabolism and renal function. *Dev Cell* 6:791-800.

Rinn, JL, M Snyder. 2005. Sexual dimorphism in mammalian gene expression. *Trends Genet* 21:298-305.

Ripley, JL. 2009. Osmoregulatory role of the paternal brood pouch for two Syngnathus species. *Comp Biochem Physiol A Mol Integr Physiol* 154:98-104.

Ripley, JL, CM Foran. 2009. Direct evidence for embryonic uptake of paternally-derived nutrients in two pipefishes (Syngnathidae: Syngnathus spp.). *J Comp Physiol B* 179:325-333.

Rivera, MC, R Jain, JE Moore, JA Lake. 1998. Genomic evidence for two functionally distinct gene classes. *Proc Natl Acad Sci U S A* 95:6239-6244.

Robison, BD, RE Drew, GK Murdoch, M Powell, KJ Rodnick et al. 2008. Sexual dimorphism in hepatic gene expression and the response to dietary carbohydrate manipulation in the zebrafish (Danio rerio). *Comp Biochem Physiol Part D Genomics Proteomics* 3:141-154.

Rosenthal, GG, XF de la Rosa Reyna, S Kazianis, MJ Stephens, DC Morizot, MJ Ryan, FJ García de León, RE Wood. 2003. Dissolution of sexual signal complexes in a hybrid zone between the swordtails Xiphophorus birchmanni and Xiphophorus malinche (Poeciliidae). *Copeia* 2003:299-307.

Sadate-Ngatchou, PI, DJ Pouchnik, MD Griswold. 2004. Identification of testosterone-regulated genes in testes of hypogonadal mice using oligonucleotide microarray. *Molecular Endocrinology* 18:422-433.

Santos, EM, P Kille, VL Workman, GC Paull, CR Tyler. 2008. Sexually dimorphic gene expression in the brains of mature zebrafish. *Comp Biochem Physiol A Mol Integr Physiol* 149:314-324.

Santos, EM, VL Workman, GC Paull, AL Filby, KJ Van Look, P Kille, CR Tyler. 2007. Molecular basis of sex and reproductive status in breeding zebrafish. *Physiol Genomics* 30:111-122.

Sayres, MAW, C Venditti, M Pagel, KD Makova. 2011. Do variations in substitution rates and male mutation bias correlate with life-history traits? a study of 32 mammalian genomes. *Evolution* 65:2800-2815.

Schein, M, Z Yang, T Mitchell-Olds, KJ Schmid. 2004. Rapid evolution of a pollen-specific oleosin-like gene family from Arabidopsis thaliana and closely related species. *Mol Biol Evol* 21:659-669.

Sievers, F, A Wilm, D Dineen, et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 7.

Singh, RS, RJ Kulathinal. 2005. Male sex drive and the masculinization of the genome. *Bioessays* 27:518-525.

Small, CM, GE Carney, Q Mo, M Vannucci, AG Jones. 2009. A microarray analysis of sex- and gonad-biased gene expression in the zebrafish: evidence for masculinization of the transcriptome. *BMC Genomics* 10:579.

Sogabe, A, Y Yanagisawa. 2007. Maintenance of pair bond during the non-reproductive season in a monogamous pipefish Corythoichthys haematopterus. *Journal of Ethology* 26:195-199.

Spence, R, G Gerlach, C Lawrence, C Smith. 2008. The behaviour and ecology of the zebrafish, Danio rerio. *Biological Reviews* 83:13-34.

Sreenivasan, R, M Cai, R Bartfai, X Wang, A Christoffels, L Orban. 2008. Transcriptomic analyses reveal novel genes with sexually dimorphic expression in the zebrafish gonad and brain. *PLoS One* 3:e1791.

Stolting, KN, AB Wilson. 2007. Male pregnancy in seahorses and pipefish: beyond the mammalian model. *Bioessays* 29:884-896.

Streelman, JT, CL Peichel, DM Parichy. 2007. Developmental genetics of adaptation in fishes: the case of novelty. *Annual Review of Ecology Evolution and Systematics* 38:655-681.

Swanson, WJ, R Nielsen, QF Yang. 2003. Pervasive adaptive evolution in mammalian fertilization proteins. *Mol Biol Evol* 20:18-20.

Swanson, WJ, VD Vacquier. 1998. Concerted evolution in an egg receptor for a rapidly evolving abalone sperm protein. *Science* 281:710-712.

Swanson, WJ, VD Vacquier. 2002a. The rapid evolution of reproductive proteins. *Nature Reviews Genetics* 3:137-144.

Swanson, WJ, VD Vacquier. 2002b. Reproductive protein evolution. *Annual Review of Ecology and Systematics* 33:161-179.

Swanson, WJ, A Wong, MF Wolfner, CF Aquadro. 2004. Evolutionary expressed sequence tag analysis of Drosophila female reproductive tracts identifies genes subjected to positive selection. *Genetics* 168:1457-1465.

Swanson, WJ, Z Yang, MF Wolfner, CF Aquadro. 2001. Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. *Proc Natl Acad Sci U S A* 98:2509-2514.

Takahashi, H. 1977. Juvenile hermaphroditism in the zebrafish, Brachydanio rerio. *Bulletin of the Faculty of Fisheries Hokkaido University* 28:57-65.

Tang, RY, A Dodd, D Lai, WC McNabb, DR Love. 2007. Validation of zebrafish (Danio rerio) reference genes for quantitative real-time RT-PCR normalization. *Acta Biochimica et Biophysica Sinica* 39:384-390.

Thornhill, R. 1983. Cryptic female choice and its implications in the scorpionfly Harpobittacus nigriceps. *American Naturalist* 122:765-788.

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

Traut, W, H Winking. 2001. Meiotic chromosomes and stages of sex chromosome evolution in fish: zebrafish, platyfish and guppy. *Chromosome Research* 9:659-672.

Tsaur, SC, CI Wu. 1997. Positive selection and the molecular evolution of a gene of male reproduction, Acp26Aa of Drosophila. *Mol Biol Evol* 14:544-549.

Turner, LM, HE Hoekstra. 2006. Adaptive evolution of fertilization proteins within a genus: variation in ZP2 and ZP3 in deer mice (Peromyscus). *Mol Biol Evol* 23:1656-1669.

Turner, LM, HE Hoekstra. 2008. Causes and consequences of the evolution of reproductive proteins. *Int J Dev Biol* 52:769-780.

Uchida, D, M Yamashita, T Kitano, T Iguchi. 2002. Oocyte apoptosis during the transition from ovary-like tissue to testes during sex differentiation of juvenile zebrafish. *J Exp Biol* 205:711-718.

Vacquier, VD, WJ Swanson, YH Lee. 1997. Positive Darwinian selection on two homologous fertilization proteins: what is the selective pressure driving their divergence? *J Mol Evol* 44:S15-S22.

Viets, BE, MA Ewert, LG Talent, CE Nelson. 1994. Sex-determining mechanisms in squamate reptiles. *Journal of Experimental Zoology* 270:45-56.

Viets, BE, A Tousignant, MA Ewert, CE Nelson, D Crews. 1993. Temperature-dependent sex determination in the leopard gecko, Eublepharis macularius. *J Exp Zool* 265:679-683.

von Hofsten, J, PE Olsson. 2005. Zebrafish sex determination and differentiation: involvement of FTZ-F1 genes. *Reprod Biol Endocrinol* 3:63.

Voolstra, C, D Tautz, P Farbrother, L Eichinger, B Harr. 2007. Contrasting evolution of expression differences in the testis between species and subspecies of the house mouse. *Genome Res* 17:42-49.

Waage, JK. 1979. Dual function of the damselfly penis: sperm removal and transfer. *Science* 203:916-918.

Wagstaff, BJ, DJ Begun. 2005. Molecular population genetics of accessory gland protein genes and testis-expressed genes in Drosophila mojavensis and D. arizonae. *Genetics* 171:1083-1101.

Walters, JR, RG Harrison. 2011. Decoupling of rapid and adaptive evolution among seminal fluid proteins in heliconius butterflies with divergent mating systems. *Evolution* 65:2855-2871.

Wang, H, ZY Gong. 1999. Characterization of two zebrafish cDNA clones encoding egg envelope proteins ZP2 and ZP3. *Biochimica Et Biophysica Acta-Gene Structure and Expression* 1446:156-160.

Wang, LK, ZX Feng, X Wang, XW Wang, XG Zhang. 2010. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* 26:136-138.

Wang, XG, R Bartfai, I Sleptsova-Freidrich, L Orban. 2007. The timing and extent of 'juvenile ovary' phase are highly variable during zebrafish testis differentiation. *Journal of Fish Biology* 70:33-44.

Wen, C, Z Zhang, W Ma, M Xu, Z Wen, J Peng. 2005. Genome-wide identification of female-enriched genes in zebrafish. *Dev Dyn* 232:171-179.

Willis, PM, MJ Ryan, GG Rosenthal. 2011. Encounter rates with conspecific males influence female mate choice in a naturally hybridizing fish. *Behavioral Ecology* 22:1234-1240.

Wilson, AB, I Ahnesjo, ACJ Vincent, A Meyer. 2003. The dynamics of male brooding, mating patterns, and sex roles in pipefishes and seahorses (family Syngnathidae). *Evolution* 57:1374-1386.

Wong, A. 2010. Testing the effects of mating system variation on rates of molecular evolution in primates. *Evolution* 64:2779-2785.

Wong, A. 2011. The molecular evolution of animal reproductive tract proteins: what have we learned from mating-system comparisons? *Int J Evol Biol* 2011:908735.

Wong, BBM, GG Rosenthal. 2006. Female disdain for swords in a swordtail fish. *American Naturalist* 167:136-140.

Wu, ZJ, RA Irizarry, R Gentleman, F Martinez-Murillo, F Spencer. 2004. A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association* 99:909-917.

Wyckoff, GJ, W Wang, CI Wu. 2000. Rapid evolution of male reproductive genes in the descent of man. *Nature* 403:304-309.

Yang, X, EE Schadt, S Wang, H Wang, AP Arnold, L Ingram-Drake, TA Drake, AJ Lusis. 2006. Tissue-specific expression and regulation of sexually dimorphic genes in mice. *Genome Res* 16:995-1004.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 13:555-556.

Yang, Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol* 15:568-573.

Yang, Z. 2006. Computational Molecular Evolution: Oxford University Press, Oxford.

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

Yang, Z, JP Bielawski. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol* 15:496-503.

Yang, Z, R Nielsen, N Goldman, AM Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155:431-449.

Yang, Z, WS Wong, R Nielsen. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. *Mol Biol Evol* 22:1107-1118.

Zeng, S, ZY Gong. 2002. Expressed sequence tag analysis of expression profiles of zebrafish testis and ovary. *Gene* 294:45-53.

Zhang, J, R Nielsen, Z Yang. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol* 22:2472-2479.

Zhang, L, WH Li. 2004. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Mol Biol Evol* 21:236-239.

Zhang, Y, D Sturgill, M Parisi, S Kumar, B Oliver. 2007. Constraint and turnover in sex-biased gene expression in the genus Drosophila. *Nature* 450:233-237.

Zhang, Z, TM Hambuch, J Parsch. 2004. Molecular evolution of sex-biased genes in Drosophila. *Mol Biol Evol* 21:2130-2139.

Zhang, Z, Y Wang, S Wang, J Liu, W Warren, M Mitreva, RB Walter. 2011. Transcriptome analysis of female and male Xiphophorus maculatus Jp 163 A. *PLoS One* 6:e18379.

APPENDIX

Additional files

Additional File 1. Annotation information and expression rankings for all male-enriched genes.

Microsoft Excel® spreadsheet containing all male-enriched GeneChip® probe sets, gene identifiers, fold change rank for each absolute expression analysis, overall mean rank, and annotation details if available.

Additional File 2. Annotation information and expression rankings for all female-enriched genes.

Microsoft Excel® spreadsheet containing all female-enriched GeneChip® probe sets, gene identifiers, fold change rank for each absolute expression analysis, overall mean rank, and annotation details if available.

Additional File 3. Genes potentially expressed differentially between male and female body.

Microsoft Excel® spreadsheet containing sex-biased genes (body) significant (FDR = 0.05) in at least one absolute expression comparison, and relevant fold change estimates.

Additional File 4. Annotation information and expression rankings for all testis-upregulated genes.

Microsoft Excel® spreadsheet containing all testis-upregulated GeneChip® probe sets, gene identifiers, fold change rank for each absolute expression analysis, overall mean rank, and annotation details if available.

Additional File 5. Annotation information and expression rankings for all ovary-upregulated genes.

Microsoft Excel® spreadsheet containing all ovary-upregulated GeneChip® probe sets, gene identifiers, fold change rank for each absolute expression analysis, overall mean rank, and annotation details if available.

Additional File 6. Real-time qPCR data.

Microsoft Excel® spreadsheet containing original qPCR expression values, relevant calculations, and statistical test details.

Additional File 7. Detailed across-study comparison of sex- and gonad-biased gene expression in zebrafish.

Microsoft Excel® spreadsheet containing the genes listed in Table 3, plus relevant expression means, standard errors, and fold change estimates for each of the four absolute expression comparisons.

Additional File 8. Zona pellucida expression data

Microsoft Excel® spreadsheet containing zona pellucida genes represented in this experiment, plus relevant expression means and standard errors for each of the four absolute expression comparisons.

VITA

Name: Clayton Matthew Small

Address: 913 Stanfield Circle,
Bryan, TX 77802

Email Address: csmall@mail.bio.tamu.edu

Education: B.S., Zoology, University of Idaho, 2003
Ph.D., Zoology, Texas A&M University, 2012