Gene Expression Regulation Evolution following Whole Genome Duplication: two comparative studies in *Xenopus* clawed frogs

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfillment of the Requirements

for the Degree

Master of Science

McMaster University

Dave W. Anderson, August 2008

Preface:

Each chapter in this thesis has been written as a separate manuscript. Data collection, analysis and written preparation for each chapter have been primarily an individual effort, with contributions in modeling, editing and result interpretation from Ben J. Evans. Chapter 2 has been submitted for publication in the Journal of Molecular Evolution, and Chapter 3 is in the final stages of preparation for submission.

Abstract:

Gene expression, and its mechanisms of regulation, remains a major area for contemporary research in evolution. With its role connecting specific gene sequences and their protein products, contributing to efforts toward understanding the specific contributions of different mechanisms of gene expression regulation is the goal of this thesis. Through two specific case studies, this thesis examines expression regulatory divergence in two different physiological contexts; the immuno-response *rag1* and *rag2* genes, and the male upregulated sex differentiation gene *dmrt1*.

Acknowledgements

There are more than a few people who deserve my utmost thanks for immeasurably valuable contributions to this work. I have tried to acknowledge specific project contributions within the body of the text, but there are several people who warrant particular recognition.

First off, my Evans labmate Iqbal Setiadi, who was both abundantly helpful when it came to lab work, and endlessly amusing when it came to averting the insanity one faces after too much lab work. Both contributions were vitally necessary to my completing this thesis.

Also, Frederic Chain, who often provided some much needed, level-headed experience to many of the problems I faced in these projects. He also deserves credit for tolerating my near-endless griping about cleaning frog cages.

My friends and colleagues in the Biology department were also of inordinate value: Carlos, Abha, Wilfred, Melanie, Steph, Wilson, Alex, Maria, Laura, Jonathan, and my illustrious committee members, Brian Golding, Rama Singh and Jon Stone. I would also like to thank the department's helpful, and oh-so-patient administrative staff: Pat, Marg, Luce, Barb, and both Kathys.

Finally, and most importantly, I must acknowledge the exemplary tutelage provided by my supervisor, Ben Evans. Upon entering graduate level study, I had neither a good understanding of the process, nor any appreciation for the importance of one's

choice of supervisor. I cannot exaggerate how lucky and fortunate I was to end up working with Ben, whose intelligence, understanding, and most importantly, boundless enthusiasm helped make pursuing research fun, rather than work. The standard of my experience has been set extraordinarily high, and I can only hope my future relationships will prove half as rewarding.

Table of Contents

Chapter 1

Chapter 2

List of Figures

List of Tables

Chapter 1

The Evolution of Gene Expression: a review of contemporary biological studies

Abstract:

Gene expression is an inextricable component of organism development and physiology. This review chapter summarizes a few prominent studies and techniques that have been used in exploring aspects of gene expression level, regulation, as well as both developmental and evolutionary changes. Working from the context of its evolutionary importance influencing development, adaptation and speciation, gene expression studies have only just begun to scratch the surface of the underlying aspects of this biological process. Techniques such as microarrays, RT-PCR and PSQ pyrosequencing are discussed, as well as the conclusions which such studies enable scientists to draw in terms of expression changes, both in terms of overall expression level and manners of regulation.

Evolution and Gene Expression

Natural selection depends on variation at the level of overall organism fitness. Differences between individuals (within a population, species, or between species) will influence mating success, and as such, there must be phenotypic variation upon which selection may act. Though most current methods of detecting selection look at gene or protein sequences, selection acts specifically upon variability in overall phenotype.

The importance of variation at the sequence level (both protein and DNA) has been well described (particularly with modern sequencing and bioinformatic techniques). Evolutionary science currently relies heavily on sequence-based analyses for phylogenetic studies, uncovering evolutionary phenomena like selection, and for investigating mutation and assessing sources of novel function. However, the link between sequence variation and phenotypic (and therefore fitness) variation has yet to be fully characterized. Potentially the most important step between gene sequence and phenotype is gene expression; in other words, the degree to (and manner in) which each gene is expressed.

It is possible to drastically affect an organism's phenotype simply via changes in gene regulatory mechanisms, and it has been postulated that evolution may be more dependant on variations in gene expression than it is on protein-level differences (King and Wilson 1975; Oleksiak et al. 2002). As such, regulating gene expression can be thought of as the step through which information encoded in the genome is

activated, thereby influencing development and determining overall phenotype. Differences in this regulation can thus manifest in different phenotypes, even between highly similar genomes.

Speciation and post-zygotic isolation

Gene expression regulation has been recognized as a vital player in evolutionary processes since Dobzhansky proposed the idea of interacting "complimentary genes" (Dobzhansky 1936). His theory suggests that incompatibilities may arise between genetically isolated populations such that an allele's deleterious effects become evident only when interacting with a different genetic background upon hybridization (Orr 1996). This provides an explanation for how post-zygotic isolation and speciation can arise without being strongly selected against, and therefore how populations between which there is reduced mating may become distinct, reproductively isolated species.

These "incompatibilities" depend on the notion that one's gene function is reliant on the proper expression of other genes elsewhere in the genome. The primary mechanism by which disparate genes interact is through transcriptional regulation (Wray et al. 2003). These regulated genetic pathways are currently the most feasible explanation for the widely prevalent complex epistatic interactions seen in all complex organisms (Johnson and Porter 2000). Thus, it is generally agreed that gene expression plays an integral role in the process of speciation via post-zygotic isolation.

Selection on Gene Expression

Given its direct connection with phenotype, gene expression is particularly vulnerable to selective pressure. However, there have thus far been few studies able to characterize the influence of selection on gene expression level, and by extension, on changes in regulatory networks. The biggest difficulty continues to be detecting a departure from neutral evolution, where the expectation is a gradual accumulation of gene expression variation with phylogenetic distance. To detect, for example, purifying selection acting on the expression of a gene, one must have a good neutral model to which empirical measurements may be compared, and the null hypothesis of 'neutral evolution' rejected. Several gene expression studies (Gilad et al. 2006a; Rifkin et al. 2005) have relied on a neutral null model for the accumulation of variation in a quantitative trait provided by Michael Lynch (Lynch 1990):

$10^{-4}V_e < V_m < 10^{-2}V_e$, where $V_w/2 \leq V_e$

V_e is the environmental variance of a trait, V_w the phenotypic variance of that trait within a population, and V_m the increase in variance of the trait per generation that is solely the result of mutation (Lynch 1990). One can predict the expected variation in expression under neutral evolution from a known phylogenetic distance and V_m. Unfortunately, V_m is difficult to measure and/or infer, even in model organisms. The best studies thus far have been conducted in *Drosophila* and *Caenorhabditis elegans*, where mutation accumulation lines have been used to estimate V_m for a

number of gene expression levels. In *Drosophila*, general conclusions find roughly 15-20% of genes exhibiting "lineage specific" expression divergence, indicating either neutral or directional selection, while most exhibit significant purifying selection (Rifkin et al. 2005; Rifkin et al. 2003). In *C. elegans*, similar work has determined the majority of genes to be under purifying selection in terms of their overall expression levels (Denver et al. 2005).

Another aim of many studies is to identify cases of positive, or directional, selection acting on gene expression level. One strategy is to identify a positive covariance between gene expression level and an environmental factor, thereby assuming that, when under selective pressure to adapt to environmental variation, gene expression level will change in turn. In *Fundulus* teleost fish, it was identified that 13 out of 58 temperature-adaptation related genes have been found to be responsive to positive selection in terms of their overall levels of expression (detected, in this case, by altering environmental temperature) (Whitehead and Crawford 2006).

In humans and primates, studies have looked at expression level variation across species, and compared it to divergence time. If there is little or no correlation between the two factors, purifying selection is the most commonly inferred selective pressure on expression. By this method, purifying selection is thought to be predominant in primates, though several families of genes have been identified that appear to be significantly influenced by directional selection, inferred for cases with significant lineage-specific divergence of expression (Gilad et al. 2006a). Many

gene expression levels are said to be under directional selection in humans, particularly genes encoding for transcription factors (Gilad et al. 2006b).

These findings have implications for speciation studies. Models have shown that directional selection on a phenotype will influence gene expression regulatory networks such that hybrid fitness between divergent species will rapidly decline, quickly reinforcing reproductive boundaries (Johnson and Porter 2000). In fact, the rapidity of this effect is maintained even if selection on divergent species is parallel (i.e. in the same direction), as the regulatory networks by which genes are expressed will still diverge in terms of their specific mechanisms, rendering the hybrid networks incompatible. This lowering of hybrid fitness is also inferred to occur under neutral or purifying selection, given sufficient phylogenetic distance, as changes in specific gene regulatory networks accumulate (Johnson and Porter 2000).

Sexual selection influences gene expression levels and regulatory networks along with natural selection in a significant and additive way. Genes involved in sexual reproduction have been shown to undergo unusually rapid rates of evolution (Kleene 2005), and it has been posited that sexual selection in divergent populations may contribute significantly to the evolution of postzygotic isolation, as a consequence of divergence in gene regulatory networks (Orr and Presgraves 2000). In *Drosophila*, the most significant expression divergence has been identified in sex-biased genes, where sexual selection may add directional pressure to gene expression levels (Lemos et al. 2005). In particular, regulatory mechanisms controlling transcription

and mRNA stability have been identified as evolving particularly rapidly in male *Drosophila* gonads (Kleene 2005).

Given the importance for speciation and the sensitivity to both natural and sexual selection, gene expression is a key player in the evolutionary development of species, and the morphological changes that occur as species lineages continue to evolve.

Present methodology for investigating gene expression regulation

Unfortunately, designing empirical studies targeting the evolution of transcription regulation networks and mechanisms along distinct lineages has proven difficult, due in part to technological impediments as well as the need to derive analytical methods that are able to distinguish between different components of gene regulation (Wray et al. 2003). Since by far the most active type of transcriptional regulation occurs at the level of transcription initiation, mRNA studies provide the best opportunity to examine facets of gene expression (Wray et al. 2003). This is because mRNA transcript abundances serve well as a surrogate for studying quantitative expression (Schadt et al. 2003). Most studies thus far have used mRNA microarrays to detect differences in expression between species, hybrids, tissues and different life stages (e.g. Chain and Evans 2008). Microarrays are particularly useful because they allow one to obtain a general survey of genome-wide expression levels (Schena et al. 1995). Unfortunately, they can be costly, and have proven difficult to use for the type of comparative studies that are required to adequately explore the contributions of different gene regulatory mechanisms due to largely unquantifiable biases

(Malone et al. 2007). (For criticisms see Chain and Evans 2008). Furthermore, microarrays often lack sufficient resolution to look at relatively small differences (i.e. on the order of 2x) in allele specific expression levels, which is particularly useful when examining expression profiles of organisms that have recently diverged (Wittkopp et al. 2006).

Real-time polymerase chain reaction (RT-PCR) provides an alternative to microarrays, and looks at gene specific expression by monitoring the degree of amplification of a gene of interest from cDNA. While this approach has many advantages, such as greater accuracy and no intrinsic biases, there are also disadvantages. First and foremost, for studies of gene expression regulatory evolution one must identify a constitutively expressed "housekeeping" gene that is expressed at reliably consistent levels (Klein 2002). This is difficult for many evolutionary studies, which use F1 cross species hybrids, in which even normally reliable genes have expression levels that are hard to predict. However, this can still be a useful technique for obtaining a physiological context of total expression levels between divergent parental species (Osawa et al. 2005).

SNP allele-quantification pyrosequencing (PSQ) provides a much rosier prospect for the sort of analysis done in comparative gene expression studies. PSQ has significant drawbacks; the narrow scale of doing gene-by-gene studies (for which one must already possess the specific cDNA sequences), and its inability to assess the total expression levels of a gene. However, the main advantage is its accurate measurement of the relative expression levels between different alleles expressed in

the same mRNA pool. Studies headed by Patricia Wittkopp and Christian Landry targeted different aspects of gene expression and regulation in *Drosophila* using pyrosequencing techniques. This approach, while limited in scope, allows greater depth in terms of examining specific mechanisms of regulation in a gene specific manner.

Mechanisms of Gene Regulation

There are many different mechanisms by which the expression of genes is regulated, each of which leaves a unique signature. There are well-documented cases of epigenetic regulation by DNA methylation and/or histone modification, which result in (usually maternal) dominant allele expression, and recessive allele repression (Jaenisch and Bird 2003). However, predominant gene expression regulation is accomplished at the level of gene transcription, which is stimulated and/or repressed by regulatory proteins, which bind to specific sequences either nearby or at considerable distance from a gene locus, influencing the activity of transcription machinery like RNA polymerase 1 (Ptashne 1986). Gene transcription regulation can be further broken down into two categories; *cis*- acting regulatory factors, which affect transcription initiation, transcription rate and/or transcript stability in an allele-specific way, and *trans*- acting factors, which interact with *cis*-regulatory sequences, but may affect expression of all gene alleles present in the cellular environment (Wittkopp et al. 2004).

Investigating the relative degree of *cis*- and *trans*- acting regulatory divergence can be accomplished via comparative analysis, particularly between relative expression

levels in parental species and the relative expression levels of species-specific alleles in F1 cross-species hybrids. Studies of this sort have been done with *Drosophila* (Landry et al. 2005; Wittkopp et al. 2004; Wittkopp et al. 2008), and applying it to another model organism is the primary goal of this thesis. These studies in *Drosophila* have found a prevalence of *cis*-acting regulatory changes between species influencing overall difference in gene expression (Wittkopp et al. 2004). Interestingly, there is often significant antagonistic regulatory divergence, with *cis*- and *trans*- regulatory changes acting in opposite directions, in a compensatory manner, resulting in similar expression between parental species. Thus, the changes only become evinced when examined in F1 hybrids. These genes are likely candidates for reinforcing post-zygotic isolation, as in most cases they exhibit hybrid misexpression of the gene (Landry et al. 2005).

The prevalence of *cis*- acting regulatory changes relative to *trans*- acting changes between divergent species has implications in terms of the mutability of these different mechanisms. Mutation studies have found that *trans*- acting regulatory factors are much more vulnerable to random changes in DNA sequences (Yvert et al. 2003), and that most induced random changes will cause *trans*- acting changes in gene expression (Morley et al. 2004). This suggests that the *trans*- mutational target size is typically much larger than the *cis*- mutational target size (Landry et al. 2007).

However, despite indications that most genes experience *trans*- altering mutations more frequently than *cis*-, interspecific studies in *Drosophila* show that *cis*- regulatory

changes accumulate preferentially over longer phylogenetic distances. This suggests that these types of changes may be favoured by natural selection (Wittkopp et al. 2008). This would be expected, as *cis*- acting changes are almost entirely gene specific (i.e. each change affects the expression of only one particular gene), while it appears that many *trans*- acting regulatory factors are "master regulators" that influence the expression of multiple gene targets, and thus will have a wider range of phenotypic effect (Morley et al. 2004). Due to the larger overall phenotypic effect of *trans*- acting regulatory changes, it is more likely that they will bring deleterious consequences along with a selectively advantageous expression change.

It is also worth noting that the most common example of regulatory divergence – *cis*-*trans*- antagonistic divergence – is predicted to happen over relatively large phylogenetic distances in genes under strong purifying selection on expression levels, as the overall expression levels may remain steady even though the specific regulatory mechanisms change.

Overall, studies investigating the prevalence and overall effects of different elements in gene regulatory networks have done much to elucidate the level of evolution linking gene expression to phenotypic selection, and extending the depth and number of these types of studies is the primary aim of this thesis.

Gene Expression and WGD: Using the *Xenopus* model

Xenopus African clawed frogs are well characterized experimental model organisms, with many resources having been already devoted to sequencing genes, constructing microarrays and deducing highly confident phylogenetic relationships between extant species (see figure 1C from Evans 2008). Additionally, there are well studied closely related outgroup species, such as *Silurana tropicalis*, whose complete genome is in the process of being sequenced (Richardson and Chapman 2003). The *Xenopus* lineage has undergone ancestral allopolyploidization by fusing two diploid genomes (see figure 1.1), and extant species express many duplicate copies of genes. Allopolyploidization, as opposed to autopolyploidization, is detected by examining duplicated genes within a species that are more closely phylogenetically related to another lower-ploidy level species than they are to each other. In *Xenopus*, the inferred allopolyploidization event is probable, but not certain, as the diploid ancestors have not been identified, and may be extinct (Evans 2007).

The ancestral gene duplication event along the *Xenopus* lineage offers a compelling opportunity for gene expression studies. Gene duplication is arguably the most important source for creating functional variation. However, having multiple copies of genes poses many challenges for an organism (especially a sexually reproducing vertebrate) as, unadulterated, this will result in increased expression of each gene, violating stoichiometric balances, and altering the finely tuned regulatory networks in the genome (Ohno 1970). The vast majority of duplicated genes are silenced in a brief evolutionary time following the duplication event (Lynch and Conery 2000). If the fitness issues surrounding changes in expression levels can be overcome (or are at least non-lethal), it is also possible for a duplicate copy of a gene to take on a new

function by changing its protein sequence (neo-functionalization) or for each copy to change such that the organism requires both copies to be present and active within a genome (sub-functionalization) (Lynch et al. 2001). It has even been shown in the *Xenopus* lineage that after allolpolyploidization selection may favour maintaining duplicates of interacting genes that are from different lineages (Evans 2007). Whichever fate a duplicate gene encounters, the key to surviving duplication events is the development of different regulatory mechanisms for each duplicate copy (Ohno 1970).

Xenopus is therefore an exciting model with which to explore issues of expression regulatory change following whole genome duplication. Many extant species will successfully form viable F1 hybrids, allowing for comparative experiments modeled from previous *cis*- vs. *trans*- regulatory studies done in *Drosophila* (Landry et al. 2005; Wittkopp et al. 2004; Wittkopp et al. 2008). For this thesis, I have specifically used the parental species *Xenopus laevis* and *X. borealis*, and their F1 hybrids, to explore the development and accumulation of *cis*- and *trans*- acting regulatory divergence between two polyploid species. Since I have worked with larger organisms, I have been able to explore regulatory divergence in different tissues of the same individuals, and to investigate the effects of tissue specific expression patterns. Furthermore, working with polyploids enables another level of analysis; within a species between paralogs (duplicate copies). Additionally, traditional analyses of divergent orthologs between different species are conducted. The specific experimental and analytical methods are detailed in chapter 2.

Scope of this Thesis

I have focused on two separate but comparable studies. Chapter 2 examines the expression divergence of *cis*- and *trans*- acting regulatory mechanisms for *rag1* and *rag2*, immuno-response genes whose protein products must interact before becoming functionally active. It is a comparative study looking at the expression divergence of a maintained duplicate (*rag1*, which has two functionally active paralogs in both XL and XB) and a singleton (*rag2*, which has undergone an ancestral gene silencing event). Furthermore, it looks at the different patterns of expression divergence detected in different tissues – a first for studies of this type, and which has important implications for whole-body expression studies. Finally, it looks at the stoichiometric sufficiency hypothesis to explain biased patterns of extant gene silencing in *rag1* by modeling the development of variation in the paralogous expression ratios of extant species, and comparing it to measured levels. Chapter 2 has been submitted for review to the Journal of Molecular Evolution.

Chapter 3 examines the expression divergence of *dmrt1*, a gene that has been implicated in the process of sexual differentiation in vertebrates, specifically in terms of testes development. Since *Xenopus* F1 cross-species hybrids violate Haldane's Rule and exhibit homogametic (male, ZZ) sterility, examining the expression profile and the divergence of regulatory mechanisms is of particular interest. The findings have implications for studies focusing on the process of sexual differentiation in amphibians, and indeed, since *dmrt1* is conserved across the sexually differentiating vertebrate lineage, may contribute to the study of this evolutionarily important

process across vertebrate species. Chapter 3 is in the final stages of preparation, and will be submitted for review in the coming weeks.

In general, this thesis includes two studies examining gene expression regulatory mechanisms, and their evolutionary divergence patterns during the important period following whole genome duplication. Additionally, it highlights an effective area for future research, pointing out ways in which this sort of study may be applied to important evolutionary questions and disease treatment challenges, as summarized in the concluding section.

References

Chain F, Evans BJ (2008) Single species microarrays in comparitive transcriptomics. unpublished

Denver DR, Morris K, Streelman JT, Kim SK, Lynch M, Thomas WK (2005) The transcriptional consequences of mutation and natural selection in *Caenorhabditis elegans*. Nature Genetics 37:544-548

Dobzhansky T (1936) Studies on hybrid sterility. II. Localization of sterility factors in Drosophila pseudoobscura hybrids. Genetics 21:113-135

Evans BJ (2007) Ancestry Influences the Fate of Duplicated Genes Millions of Years After Polyploidization of Clawed Frogs (Xenopus). Genetics 176:1119-1130

Evans BJ (2008) Genome evolution and speciation genetics of clawed frogs (*Xenopus* and *Silurana*). Frontiers in Bioscience 13:4687-4706

Gilad Y, Oshlack A, Rifkin SA (2006a) Natural selection on gene expression. Trends in Genetics 22:456-461

Gilad Y, Oshlack A, Smyth GK, Speed TP, White KP (2006b) Expression Profiling in primates reveals a rapid evolution of human transcription factors. Nature 440:242-245

Jaenisch R, Bird A (2003) Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nature Genetics 33:245-254

Johnson NA, Porter AH (2000) Rapid Speciation via Parallel, Directional Selection on Regulatory Genetic Pathways. Journal of Theoretical Biology 205:527-542

King M-C, Wilson AC (1975) Evolution at Two Levels in Humans and Chimpanzees. Science 188:107-116

Kleene KC (2005) Sexual selection, genetic conflict, selfish genes, and the atypical patterns of gene expression in spermatogenic cells. Developmental Biology 277:16-26

Klein D (2002) Quantification using real-time PCR technology: applications and limitations. Trends in Molecular Medicine 8:257-260

Landry CR, Lemos B, Rifkin SA, Dickinson WJ, Hartl DL (2007) Genetic Properties Influencing the Evolvability of Gene Expression. Sciencexpress

Landry CR, Wittkopp PJ, Taubes CH, Ranz JM, Clark AG, Hartl DL (2005) Compensatory *cis-trans* Evolution and the Dysregulation of Gene Expression in Interspecific Hybrids of Drosophila. Genetics 171:1813-1822

Lemos B, Meiklejogn CD, Caceres M, Hartl DL (2005) Rates of divergence in gene expression profiles of primates, mice and flies: Stabilizing selection and variability among functional categories. Evolution 59:126-137

Lynch M (1990) The Rate of Morphological Evolution in Mammals from the Standpoint of the Neutral Expectation. The American Naturalist 136:727-741

Lynch M, Conery JS (2000) The Evolutionary Fate and Consequences of Duplicate Genes. Science 290:1151-1155

Lynch M, O'Hely M, Walsh B, Force A (2001) The Probability of Preservation of a Newly Arisen Gene Duplicate. Genetics 159:1789-1804

Malone JH, Chrzanowski TH, Michalak P (2007) Sterility and Gene Expression in Hybrid Males of *Xenopus laevis* and *X. muelleri*. PLOS One 2:1-13

Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG (2004) Genetic analysis of genome-wide variation in human gene expression. Nature 430:743-747

Ohno S (1970) Evolution by Gene Duplication. Springer-Verlag, New York

Oleksiak MF, Churchill GA, Crawford DL (2002) Variation in gene expression within and among natural populations. Nature Genetics 32:261-266

Orr HA (1996) Dobzhansky, Bateson, and the Genetics of Speciation. Genetics 144:1331-1335

Orr HA, Presgraves DC (2000) Speciation by postzygotic isolation: forces, genes molecules. BioEssays 22:1085-1094

Osawa N, Oshima Y, Nakamura M (2005) Molecular cloning of Dmrt1 and its expression in the gonad of Xenopus. Zool. Sci. 22:681-687

Ptashne M (1986) Gene regulation by proteins acting nearby and at a distance. Nature 322:697-701

Richardson PM, Chapman J (2003) The Xenopus Tropicalis Genome Project. Current Genomics 4:645-652

Rifkin SA, Houle D, Kim J, White KP (2005) A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression. Nature 438:220-223

Rifkin SA, Kim J, White KP (2003) Evolution of gene expression in the *Drosophila melanogaster* subgroup. Nature Genetics 33:138-144

Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G, Linsley PS, Mao M, Stoughton RB, Friend SH (2003) Genetics of gene expression surveyed in maize, mouse and man. Nature 422:297-302

Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. Science 270:467-470

Whitehead A, Crawford DL (2006) Netral and adaptive variation in gene expression. Proceedings of the National Academy of Sciences 103:5425-5430

Wittkopp PJ, Haerum BK, Clark AG (2004) Evolutionary Changes in *cis* and *trans* gene regulation. Nature 430

Wittkopp PJ, Haerum BK, Clark AG (2006) Parent-of-Origin Effects on mRNA Expression in *Drosophila* Not Caused by Genomic Imprinting. Genetics 173:1817-1821

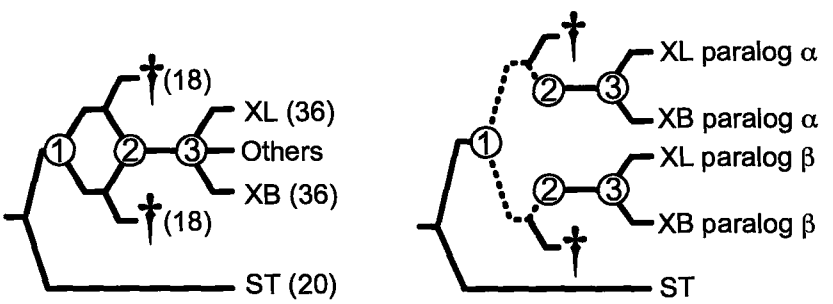
Wittkopp PJ, Haerum BK, Clark AG (2008) Regulatory changes underlying expression differences within and between *Drosophila* species. Nature Genetics 40:346-350

Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA (2003) The Evolution of Transcriptional Regulation in Eukaryotes. Mol. Biol. Evol. 20:1377-1419

Yvert G, Brem RB, Whittle J, Akey JM, Foss E, Smith EN, Mackelprang R, Kruglyak L (2003) *Trans*-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. Nature Genetics 35:57-64

Figure 1.1:

An example of a phylogeny and genealogy, illustrating relationships between *X. laevis*, *X. borealis*, and *S. tropicalis*, and duplicate paralogs (α and β). Dotted lines represent diploid species or genes, red lines are tetraploid species or genes immediately after duplication, and blue lines are tetraploid species of genes at a later stage. Taken from Chain and Evans (2006).



Anderson and Evans, Fig. 4

Title: Regulatory evolution of a duplicated heterodimer across species and tissues of allopolyploid clawed frogs (*Xenopus*).

Authors: Dave Anderson and Ben J. Evans[§]

Affiliation: Center for Environmental Genomics, Department of Biology,
Life Sciences Building Room 328 McMaster University,
1280 Main Street West
Hamilton, ON, Canada L8S 4K1

[§]Corresponding author

Correspondence:
Ben Evans
Department of Biology
1280 Main Street West, LSB room 328
McMaster University
Hamilton, ON L8S4K1
Canada
email: evansb@mcmaster.ca
phone: 905-525-9140 x 26973
fax: 905-522-6066

Abstract

Changes in gene expression contribute to reproductive isolation of species, adaptation, and development, and may impact the genetic fate of duplicated genes. African clawed frogs (genus *Xenopus*) offer a useful model for examining regulatory evolution, particularly after gene duplication, because species in this genus are polyploid. Additionally, these species can produce viable hybrids, and expression divergence between co-expressed species-specific alleles in hybrids can be attributed exclusively to *cis*- acting mechanisms. Here we have explored expression divergence of a duplicated heterodimer composed of the recombination activating genes 1 and 2 (RAG1 and RAG2). Previous work identified a phylogenetically biased pattern of pseudogenization of RAG1 paralogs wherein one paralog – RAG1β – was more likely to become a pseudogene than the other one – RAG1α. In this study we show that ancestral expression divergence between paralogs could account for this. Furthermore, using comparative data we demonstrate that the mechanisms and magnitude of regulatory divergence between species varies significantly across tissue types. These results have implications for our understanding of variables that influence pseudogenization of paralogs generated by whole genome duplication, and for interpretation of the relative contributions of *cis*- versus *trans*- mechanisms to expression divergence at the cellular level.

Key Words

cis- and *trans*- regulation, expression divergence, gene duplication, allopolyploidization, pseudogenization.

Introduction

The impact of natural selection on regulatory evolution is best understood through comparative analyses – either between different genetic conditions (such as species, populations, hybrids, strains, or tissues) or between different types of genetic information (such as gene expression level, expression divergence, DNA sequence divergence, breadth of expression, or protein-protein interactions). To this end, we have studied expression divergence of transcripts that encode recombination activating genes 1 and 2 protein (RAG1 and RAG2) in African clawed frogs, genus *Xenopus*, in different species and in different tissues. These proteins form a heterodimer that is required to carry out V(D)J recombination – the somatic shuffling of "variable", "diversity", and "joining" cassettes of DNA that makes possible the adaptive immune response (Bassing et al. 2002; Spicuglia et al. 2006). Tetraploid *Xenopus* species have 36 chromosomes and are collectively derived from a single tetraploid ancestor. This tetraploid ancestor probably originated via allopolyploidization, a process that duplicated the entire genome and generated two paralogs (α and β) of all nuclear genes (Evans 2007; Evans et al. 2005). Octoploid and dodecaploid species also exist. Octoploid *Xenopus* species have 72 chromosomes and were formed on multiple occasions through allopolyploidization of two tetraploid ancestors (Evans 2007; Evans et al. 2005). In the absence of gene deletion or pseudogenization, octoploid species are expected to have four paralogs of all nuclear genes (α_1, α_2, β_1, and β_2) because an α paralog and a β paralog was inherited from each tetraploid ancestor.

In different species of *Xenopus*, multiple paralogs of RAG1 experienced pseudogenization in a way that appears to be biased by ancestry (Evans 2007). Between seven and 16 independent instances of pseudogenization occurred in RAG1β whereas only two instances occurred in RAG1α (Evans 2007). In contrast, RAG2α was deleted from the genome prior to the diversification of extant *Xenopus* species, so their genomes now encode only RAG2β (Evans 2007). One explanation for non-random pseudogenization of RAG1β is that low expression of this paralog made it insufficient to carry out all necessary activities on its own after whole genome duplication (WGD), whereas expression of RAG1α is both sufficient and therefore necessary after WGD. We call this the "stoichiometric sufficiency of RAG1α" (SSRα) hypothesis. To test this, we estimated ancestral expression ratios of RAG1 paralogs (α/β) based on expression ratios of eight extant species, including four tetraploid species (*X. borealis*, *X. muelleri*, X. *laevis*, and *X. gilli*) and four octoploid species (*X. itombwensis*, *X. andrei*, *X. boumbaensis*, and *X. amieti*). The RAG1β_2 paralog of three of these octoploids (*X. andrei*, *X. boumbaensis*, and *X. amieti*) was probably deleted, but RAG1β_1 is still expressed even though in each of these octoploid species this paralog is degenerate at the coding level due to nonsense and missense mutations. In *X. itombwensis*, RAG1β_1 and RAG1β_2 are both present in the genome but both are degenerate at the coding level. All eight of these species, however, still express at least one RAG1α paralog and at least one RAG1β paralog.

Mechanisms of expression divergence. With an aim of better understanding regulatory evolution of this duplicated heterodimer, we also examined mechanisms

of expression divergence in RAG1 and RAG2, focusing our analysis on two tetraploid species – *X. laevis* (XL) and *X. borealis* (XB), and hybrids generated by a cross between these species (H_{XLXB}). XL and XB diverged from a common ancestor roughly 21 – 41 million years ago (Fig. 3A,B; Chain and Evans 2006; Evans 2007; Evans et al. 2005; Evans et al. 2004).

On a genetic level, expression divergence is achieved by mechanisms that operate in *cis*- that affect transcription in an allele specific manner, and by mechanisms operating in *trans*- that affect regulation of both alleles via direct or indirect interaction with *cis*- regulatory elements. *Cis*- divergence is indicated if alleles are divergently expressed in the same cell. *Trans*- divergence is indicated if alleles are differently expressed in different cells, but not when expressed in the same cell. These mechanisms drive expression divergence between alleles of a single gene, between alleles of genes generated by duplication (paralogs), and between alleles of genes generated by speciation (orthologs).

Divergent expression of orthologs could be a consequence either of *cis*- or *trans*- regulatory divergence, or both (Fig. 1). Divergent expression of paralogs of a gene that are co-expressed within a cell necessarily implies *cis*- divergence (Fig. 1). However, if each paralog of a gene is expressed in a different tissue type within the same individual, then divergent expression between them could also be achieved by paralog-specific *trans*- factors that interact with unique *cis*- regulatory sequences on each one. The same is true of paralogs of a gene that are expressed in different species – expression divergence between paralog α in species 1 and paralog β in

species 2 could be achieved by *cis*- or *trans*- divergence, or both (Fig. 1; Evans 2008).

Quantification of *cis*- and *trans*- divergence. Expression analysis of interspecies hybrids provides a way to tease apart the contributions of *cis*- and *trans*-acting mechanisms to expression divergence between orthologous alleles (Landry et al. 2005; Wittkopp et al. 2004). In hybrids, *trans*-acting factors from both species can interact with the *cis*-acting factors of species-specific alleles; so their differential expression in hybrids can be attributed exclusively to *cis*- divergence (Fig. 2). Under a scenario of exclusively *cis*- divergence, *trans*-acting factors expressed in a hybrid are the same, even though they are encoded by parental alleles from different species. This is true even though the overall expression level of these transcription factors (in terms of mRNA per cell) may differ between parental and hybrid individuals. Alternatively, divergent expression between species could be only due to *trans*-acting mechanisms (Fig. 2). Under this scenario, both species-specific alleles would be expressed at the same level in hybrids, even though these alleles are expressed at different levels in each parental species.

Other scenarios involving both *cis*- and *trans*- divergence between orthologs are possible. If *cis*- and *trans*- mechanisms both act synergistically to upregulate expression in one species compared to another, the level of expression of species-specific alleles should be more similar in hybrids than in the parental species (Fig. 2). Alternatively, if *cis*- and *trans*- acting factors influence expression divergence in opposite (antagonistic) directions, the expression level of species-specific alleles should be more similar in the parental species than in hybrids (Fig. 2; Landry et al.

2005; Wittkoop et al. 2004; Yan et al. 2002). This is because species-specific *cis-trans-* antagonism is disrupted in hybrids because *trans-* acting factors of both species interact with *cis-* elements of both parental alleles.

Predictions. With respect to the mechanisms of regulatory evolution of the RAG1-RAG2 heterodimer in *Xenopus*, we had four predictions. First, we predicted that the magnitude of *cis-* divergence between orthologs would be tissue-specific (Prediction 1). We reasoned that *cis-* and *trans-* divergence between orthologs should be detectable only to the extent that *trans-* acting factors that interact with divergent *cis-* regulatory elements are expressed in a given tissue. Second, following the same reasoning, we expected the magnitude of expression divergence between paralogs that are co-expressed in different tissues to vary among tissues (Prediction 2). We expected this because *trans-* acting factors that control RAG1 expression could be expressed at different levels in different tissues and/or because different suites of *trans-* acting factors may be expressed in different tissues. Third, because these proteins interact, we predicted that the hierarchy of expression across tissue types would be similar in RAG1α, RAG1β, and RAG2β (Prediction 3). And fourth, because RAG1 but not RAG2β has a co-expressed paralog (Evans 2007), we expected that expression divergence between RAG1 orthologs would be greater than between RAG2β orthologs (Prediction 4).

Methods

Paralogous expression divergence. We used the program BayesTraits (Organ et al. 2007) to analyze paralogous expression divergence across species in a phylogenetic context. This program reconstructs a posterior distribution of ancestral states at internal nodes using a Markov Chain Monte Carlo (MCMC) approach (Pagel 1999). The underlying model used in the reconstruction is one of Brownian motion in which the variance between an ancestral and descendant trait is expected to increase over time, even though the mean value of the descendant trait is expected to be equal to the ancestral trait. A first step in this analysis is to generate a posterior distribution of the variance parameter (α) given the topology and the observed data, which in our case was the $\log_2$-transformed paralogous expression ratio in bone marrow of eight species. We generated this distribution from 5,000,000 iterations, sampled every 1,000 iterations, after a burnin of 50,000 iterations. Another MCMC chain was then performed using a chain length of 250,000,000 iterations, sampling every 5,000 iterations, and using the same burnin, but this time drawing from the posterior distribution of α. We report the mean posterior probability of the null SSRα hypothesis – that there is either no ancestral bias ($\log_2$-transformed ratio = 0) or that there is ancestral upregulation of RAG1β ($\log_2$-transformed ratio < 0), based on ten independent MCMC chains. We did not implement topology scaling parameters in this analysis, such as κ, δ, and λ (Pagel 1997; Pagel 1999; Venditti et al. 2006) because of concerns that estimates of these parameter values may be inaccurate on a small phylogeny (A. Meade, personal communication).

The reticulating evolutionary history of clawed frogs poses unique challenges to comparative analyses because most approaches for analyzing continuous

characters within a phylogenetic context rely on a bifurcating evolutionary relationship among the species being compared. However, because sequences are now available for essentially all paralogs of all species from RAG1 and RAG2 (Evans 2007; Evans et al. 2005), we were able to construct a bifurcating phenogram based on the average pairwise genetic distance between each orthologous pair of paralogs, and then use this topology for comparative analysis. To accomplish this, an ultrametric phylogeny was estimated with a strict molecular clock enforced based on combined data from RAG1 and RAG2 using the program MrBayes version 3.1.2 (Huelsenbeck and Ronquist 2001). For each pairwise comparison between the eight species in our analysis, we calculated and averaged the patristic distances between each pair of orthologs to generate pairwise genetic distances between each pair of species (Fig. 3C). In these comparisons, averages were performed in a hierarchical way to ensure that each of the individual pairwise distance between orthologs contributed appropriately to the aggregate pairwise distance between the species, including in comparisons involving octoploid species in which one RAG1β paralog was deleted. Based on previously published phylogenetic relationships (Evans 2007; Evans et al. 2005), we substituted the genetic distances between evolutionarily equivalent comparisons with the mean genetic distance in each evolutionarily equivalent comparison, for example between pairs of sister species. The program PHYLIP version 3.6 (Felsenstein 2005) was then used to transform the resulting triangular matrix into a phenogram using the UPGMA algorithm. While this phenogram is topologically inaccurate because it does not reticulate, it is nonetheless a reasonable depiction of genetic distance between species in terms of

branch lengths and well-supported evolutionary relationships between these species (Suppl. Fig. 1).

Standardization of paralogous expression ratios. Because there could be biases in the efficiency of PCR amplification between paralogs, and also because there is an unequal number of RAG1α and RAG1β paralogs in some octoploid species, we standardized the paralogous expression ratios by subtracting the $\log_2$-transformed paralogous ratio that was obtained from gDNA. Biological replicates of these species were not available, but at least two technical replicates were performed for each one. To complement these assays, we also inferred with three biological replicates the paralogous expression ratios for XL and XB from the tissue mixes. This second analysis required additional measures that are described below to account for differences in the amount of tissue from each species (XL or XB) in the parental mix.

Orthologous expression divergence. In order to quantify the relative contributions of *cis*- and *trans*- acting factors in orthologous expression divergence, we compared the expression ratios of species-specific alleles in XL and XB to the expression ratio of these alleles in XL x XB hybrid individuals (Fig. 2). For each locus we tested whether (a) the $\log_2$-transformed parental expression ratio was significantly different from the $\log_2$-transformed hybrid expression ratio, which would be consistent with *trans*- divergence, and whether (b) the $\log_2$-transformed expression ratio in hybrids was significantly different from zero, which would be consistent with *cis*- divergence (Fig. 2; Landry et al. 2005; Wittkopp et al. 2004).

Laboratory procedures for quantification of parental and hybrid expression ratios involve a few ingenious techniques developed by others (Landry et al. 2005; Wittkopp et al. 2004). A first challenge is to directly compare the expression intensity of alleles in each parental species while accounting for differences in the total number of cells from each parental species that is used in each RNA extraction. This is accomplished by homogenizing tissue samples from both parental species (to generate a "parental mix"), and then performing both DNA and RNA extractions on this mixture. The parental expression ratio is estimated by measuring expression of alleles from each parental species in the parental mix based on species-specific and paralog-specific single nucleotide polymorphisms (SNPs), and then standardizing this expression ratio by (i) the proportion of tissue from each parental species that was in the parental mix, and (ii) the PCR amplification bias for each species-specific allele. This first proportion (i) is estimated from the ratio of each type of parental DNA in the genomic DNA extraction from the parental mix. The second proportion (ii) is estimated from the ratio of amplified parental alleles from a genomic DNA extraction from F1 hybrids (which have equal allelic concentrations from each parental species).

Parental mixes were generated by combining similarly sized tissue samples from each of the parental species, adding a small quantity of RNAse free water (~100 μl), and homogenizing the mixture by passing it through a 20.5 gauge needle or using a tissue homogenizer (PRO Scientific Inc.). RNA was extracted and converted into cDNA using the RNeasy extraction kit (Qiagen) and the cDNA synthesis Omniscript

Reverse Transcriptase kit (Qiagen). Genomic DNA was extracted using the QLAmp kit (Qiagen). The proportion of DNA or RNA from each parental species that was in each RNA and DNA extractions was quantified using a Biotage PSQ96 pyrosequencer (Kruckeberg and Thibodeau 2004; Nilsson and Johansson 2004). After amplification, species-specific alleles of RAG2β were quantified with one pyrosequencing primer whereas amplified species-specific alleles of RAG1α, RAG1β, and RAG1 paralogs of XL and XB were quantified with three different pyrosequencing primers. Primers and SNP information is in Supplementary Material 1 and are based on previously published sequences of RAG1 and RAG2 (Evans 2007).

We performed these assays on RAG1 and RAG2 alleles expressed in bone marrow, brain, heart, and liver. For each tissue type, ratios were estimated in F1 hybrids generated from a cross between an XL female and an XB male (2 females and 2 males) or the reciprocal cross (0 females and 2 males), and compared to ratios the corresponding expression ratios in 7 parental mixes (3 parental mixes between XL and XB females and 4 parental mixes between XL and XB males). For some treatments we performed replicate DNA or RNA extractions on an individual and/or replicate PCR reactions on DNA or RNA.

Standardization of orthologous expression ratios. We standardized expression ratios of species-specific alleles in the parental mixes and hybrids using the method of (Landry et al. 2005). We fit a regression line between $\log_2$-transformed parental genomic ratios from the three pyrosequencing assays (RAG2β, RAG1α, RAG1β),

and the median log_2-transformed hybrid genomic ratio that was calculated for each primer from at least six biological replicates. For each parental tissue mix, log_2-transformed expression ratios were standardized by subtracting the fitted estimate of parental DNA for each primer pair on median ratio hybrid gDNA amplified with that primer pair. This accommodates PCR bias, pyrosequencing bias, and differences in concentration of each parental tissue in the parental mix. Log_2-transformed hybrid ratios were standardized by subtracting the median ratio from hybrid gDNA amplified with that primer pair, which accounts for PCR bias and pyrosequencing bias.

Expression hierarchy. The hierarchy of expression levels among tissue types within each sex was evaluated with a series of "tissue crosses" comprised of one tissue type from XL with another tissue type from XB. We analyzed a mixture of XL liver with XB heart, XL heart with XB brain, and XL brain with XB bone marrow, and we did these comparisons separately for each sex. Log_2-transformed tissue cross expression ratios were standardized by subtracting the corresponding log_2-transformed gDNA ratio of the tissue crosses. The resulting standardized ratios were then used to calculate intraspecific expression ratios between tissue types. For example, the XL_{liver}/ XB_{heart} expression ratio was divided by the parental XL_{liver}/ XB_{liver} expression ratio to obtain the expression ratio of XB_{liver}/ XB_{heart}. This allowed us to estimate the expression hierarchy across four tissue types for each gene within each parental species and within each sex.

Statistical analysis. We used the methods of Landry et al. (Landry et al. 2005) to test whether the expression ratios were significantly different from each other, and to test whether expression ratios were significantly different from zero. We used

restricted maximum likelihood (REML) to independently estimate variance parameters for each treatment being compared. Analyses were carried out using the "proc mixed" in SAS version 9.1.3 (SAS Institute) with modified scripts that were graciously provided by Patricia Wittkopp. Similar to (Landry et al. 2005), this approach was justified to avoid bias because the variance in expression ratio among replicates was higher in the parental mixes than in the hybrids. Student's *t*-tests were computed within the mixed procedure and locus-level significance was interpreted after sequential Bonferroni correction for multiple tests (Rice 1989). Calculations included a random effect estimated from the variance between PCR replicates (same extraction but different PCR reaction), which were nested in technical replicates (separate cDNA extractions), and within biological replicates (different individuals). Results from these analyses were very similar to those recovered from regular Student's t-tests.

Because an inference of *cis*- only divergence or *trans*- only divergence is made on the basis of rejection of one hypothesis but inability to reject another (Fig. 2), these inferences suffer from Type II error. Neither inference, therefore, is statistically inconsistent with an inference of *cis*- and *trans*- divergence, which is based on rejection of two null hypotheses (indicating that $P \neq H$ and $H \neq 0$; Fig. 2). Application of the Bonferroni correction therefore makes the detection of antagonistic (or synergistic) *cis*- and *trans*- divergence more conservative. We also applied the Bonferroni correction to comparisons between expression ratios across tissue types. This makes the test for tissue-specific orthologous and paralogous *cis*- divergence more conservative.

Results

Non-random pseudogenization of RAG1. In bone marrow, RAG1α was expressed more highly than RAG1β in 7 out of 8 species that we examined. This was true even in octoploid species in which the RAG1β_1 paralog was deleted (*X. andrei*, *X. boumbaensis*, *X. andrei*) – in other words, in these species, even though there are two copies of RAG1α and one copy of RAG1β, expression of RAG1α_1 plus RAG1α_2 was more than twice as high as the expression of RAG1β_2. Additionally, in some other tissues this bias was extreme – expression of RAG1α in *X. muelleri* brain, for instance, was almost 10 fold higher than RAG1β (Table 1). Ancestral reconstruction did not reject the null hypothesis that the $\log_2$-transformed paralogous expression ratio of bone marrow was less than or equal to zero in the most recent common ancestor (MRCA) of all *Xenopus* (node 3, Fig. 3A,B; P = 0.0726) but it did reject the null hypothesis that this ratio was less than or equal to zero in the MRCA of a subset of these species (node 4, Fig. 3A,B; P = 0.0497).

Unfortunately, we were unable to obtain biological replicates from most of these species for which we analyzed paralogous expression ratios. However, we were able to infer species-specific paralogous expression ratios in XL and XB from the parental mixes, and these $\log_2$-transformed ratios did not depart significantly from zero in XL or XB bone marrow, brain, heart, or liver (data not shown). When the analyses were repeated with the expression ratio of these species set to zero, the null hypothesis that the ancestral expression ratio was less than or equal to zero was again rejected for the MRCA of the a subset of these species (node 4, Fig. 3A,B; P =

0.0492) and not rejected for the MRCA of all *Xenopus* (node 3, Fig. 3A,B; P = 0.0826). Overall, we interpret these results as providing support the SSRα hypothesis – at least for a subset of these species.

Mechanisms for expression divergence. We explored mechanisms of expression divergence by comparing expression ratios of species-specific alleles in parental and hybrid individuals, using the approach of (Landry et al. 2005; Wittkopp et al. 2004). Tests for parent of origin effects and sex bias were not significant within each tissue type for orthologous ratios of RAG1α, RAG1β, or in RAG2β or for paralogous ratios of RAG1 of XL or of XB (Suppl. Table 1). For subsequent comparisons we therefore treated these data as biological replicates, and did not perform a separate test for males and females or for each type of hybrid cross (XL or XB mother).

RAG2β. Upregulation of XL RAG2β due to *cis*- divergence was detected in two tissue types ($H \neq 0$ in bone marrow, heart, and almost significant in brain; Table 2). Parsimonious explanations for this are that these divergent *cis*- elements were (a) driven by the same *trans*- acting factors in each of these tissues or (b) driven by different tissue-specific *trans*- acting factors that have overlapping *cis*- targets. In bone marrow and heart, antagonistic *trans*- upregulation of XB RAG2β was also detected and as a result, no significant difference in expression ratios of RAG2β was detected between tissue types in the parental species (Table 3).

The expression ratios of parental alleles of RAG2β in hybrids were significantly different in multiple pairwise comparisons between tissues, even after correction for

multiple tests, indicating variation between tissue types in the magnitude of *cis*-divergence (Table 3). For example, the XL allele was much more highly expressed compared to the XB allele in hybrid heart (log_2-transformed hybrid ratio = 1.1566) than in hybrid bone marrow (log_2-transformed hybrid ratio = 0.0825, Table 2) and this difference is significant (P = 0.0001; Table 3). This significant difference is consistent with Prediction 1 – that the magnitude of *cis*- divergence can be tissue specific as a consequence of tissue specific *trans*- factors.

RAG1α and RAG1β. In most tissues, orthologous expression of XL and XB alleles of each RAG1 paralog was generally not significantly different in the parental species (null hypothesis of P = 0 not rejected; Table 2). An exception to this is expression divergence of RAG1α in liver, which exhibits antagonism between *cis*- and *trans*- divergence: *cis*- divergence upregulates XL and *trans*- divergence upregulates XB. As a result of these antagonistic mechanisms, the parental expression ratio of RAG1α in liver is not significantly different from the other tissues (data not shown). In hybrid liver the expression ratio of XL RAG1α and XB RAG1α is higher than in hybrid bone marrow and brain (Tables 2, 3), which is also consistent with tissue specific *cis*- divergence (Prediction 1), but this difference is not significant after correction for multiple tests.

Cis- divergence in bone marrow causes upregulation of XL RAG1β (Table 2). This could also include an antagonistic *trans*- contribution – which could explain why the RAG1β expression ratio is not significantly different in bone marrow from the other tissues – but evidence for *trans*- divergence of RAG1β was not statistically

significant in bone marrow (Table 2). As a consequence of tissue-specific *cis*-divergence, upregulation of XL RAG1β compared to XB RAG1β is significantly higher in hybrid bone marrow than in hybrid heart or hybrid liver (Table 3). Similar to our observations in RAG2β, this is also consistent with Prediction 1.

RAG1 paralogs of XL and XB also exhibit tissue specific signature of *cis*-divergence, which is consistent with Prediction 2 (Table 4). Tissue specificity of paralogous *cis*- divergence is evinced, for example, by comparing the expression ratios of XL paralogs in hybrid liver to hybrid bone marrow. In liver, XL RAG1α is significantly upregulated compared to XL RAG1β (log_2-transformed hybrid ratio = 0.4625) but in bone marrow it XB RAG1β is upregulated (log_2-transformed hybrid ratio = -0.8587). This difference in paralogous expression between tissue types is significant (P = 0.0010, Table 4).

Expression hierarchy; orthologous divergence in singletons and duplicates. Consistent with Prediction 3, the hierarchy of expression levels in different tissues of RAG2β, RAG1α, and RAG1β is similar within and between each species (Table 5). No significant difference was detected between these genes in the hierarchy of expression across tissues and there was substantial qualitative consistency among these hierarchies. Expression was generally highest in bone marrow. The next highest expression was in brain, followed by heart and then liver, except in RAG1β where expression was higher in liver than heart. These results are consistent with another study (Greenhalgh et al. 1993) that found higher expression of RAG1 and RAG2 in adult bone marrow than in liver.

Contrary to Prediction 4, expression divergence between orthologs of RAG1 was not of substantially higher magnitude than RAG2β. The absolute value of the $\log_2$-transformed expression divergence across parental tissue types was not significantly higher in orthologs of RAG1α or of RAG1β than in orthologs of RAG2β (Suppl. Table. 2). In fact, expression divergence in hybrids – which reflect *cis*- only divergence – was greater in the singleton (RAG2β) than in either co-expressed paralogs of RAG1 in bone marrow and heart.

Discussion

In African clawed frogs of the genus *Xenopus*, genome duplication gave rise to duplicated copies (paralogs) of the RAG1 gene; these paralogs then degraded in a seemingly biased fashion wherein one set of closely related paralogs tended to become pseudogenes more frequently than another (Evans 2007; Evans et al. 2005). One explanation for this, which we call the SSRα hypothesis, posits that expression divergence between RAG1 paralogs favors the retention of the one that was upregulated (RAG1α) in an early ancestor. In multiple tissues and species, this hypothesis is supported by higher expression of RAG1α (Table 1) and a reconstructed ancestral expression ratio in bone marrow is also consistent with ancestral bias. We note, however, that there is a "chicken and egg" problem in the testing of the SSRα hypothesis because it is not clear whether low expression of RAG1β in these octoploids is a cause or a consequence of pseudogenization of some RAG1β paralogs in octoploids. *X. boumbaensis*, *X. andrei*, *X. amieti*, and *X. itombwensis* all have nonsense or frameshift mutations in $RAG1\beta_1$; *X. itombwensis*

also has a frameshift mutation in RAG1β_2 whereas these other octoploids appear to be missing this paralog, probably as a consequence of a single ancestral deletion (Evans 2007). Additionally, one *X. borealis* individual was identified that was heterozygous for a nonsense mutation in RAG1β (Evans 2007). We therefore can not rule out other explanations for biased RAG1 pseudogenization such as that (a) non-lethal incompatibilities exist between RAG1β and cofactors in the other subgenome (the "α" subgenome) that increase the probability of pseudogenization of this paralog, or that (b) advantageous interactions between RAG1α and RAG2β (or some other gene) increase the probability of persistence of this paralog compared to RAG1β. These other possibilities make opposite inferences about the nature of natural selection acting on paralogs of RAG1, i.e. that either negative selection on RAG1β or positive selection on RAG1α contributed to the non-random pseudogenization of RAG1β.

Because gene duplication changes gene dosage, these two phenomena have intertwined implications for development and evolution. An increased understanding of the relative contributions of *cis*- and *trans*- factors in expression divergence, particularly after gene or genome duplication, would offer perspective on fundamental genetic mechanisms that underlie adaptation, reproductive isolation, and phenotypic plasticity. However, conclusions regarding the relative importance of *cis*- and *trans*- mechanisms in expression divergence vary among studies, among genes, and perhaps among species (Cowles et al. 2002; Landry et al. 2005; Wittkoop et al. 2004). Expression analyses in humans, yeast, flies, nematodes, and *Eucalyptus* suggest an important role for *trans*-acting factors (Brem et al. 2002; Kirst

et al. 2005; Morley et al. 2004; Yvert et al. 2003), whereas other reports on humans, flies, and mice recover an important role for *cis*-acting factors (Cowles et al. 2002; Wittkoop et al. 2004; Wittkopp et al. 2008; Yan et al. 2002). This issue has important implications for our understanding of the genetic architecture and evolution of gene regulation. For example, if *cis*- divergence is less common than *trans*- divergence, *trans*- regulatory factors are either a comparatively larger mutational target or they are under comparatively less severe purifying selection (or more extensive positive or diversifying selection) than *cis*- regulatory elements. Similar to some of these previous studies, our results indicate that *cis*- and *trans*- mechanism both sculpted expression divergence of the RAG1-RAG2 heterodimer in clawed frogs. Many orthologous expression ratios of RAG1 and RAG2 are not significantly diverged between these parental species, but comparison to hybrid expression uncovers antagonistic interactions between *cis*- and *trans*- divergence, including *cis*- upregulation of XL alleles coupled with *trans*- upregulation of XB alleles in some tissues (Table 2).

Selective pressures on regulatory evolution. Purifying selection can lead to a conserved level of expression even when regulatory sequences have changed considerably (Ludwig et al. 2000). This can happen if compensatory changes occur in different components of the regulatory machinery. In the nematode *Caenorhabditis elegans*, comparison of mutation accumulation lines with a low effective population size (N_e) had more expression divergence than divergent natural isolates with a larger N_e (Denver et al. 2005). This suggests purifying selection on gene regulation in the natural isolates. Comparison of orthologous divergence in fruit flies, primates, and rodents suggests that the overall level of expression is

conserved in many genes, which is also consistent with purifying selection (Jordan et al. 2005; Lemos et al. 2005). Intraspecific polymorphism in expression in *Drosophila* appears to be largely attributable to *trans*- mechanisms whereas intraspecific expression divergence appears to be primarily a consequence of *cis*- mechanisms, suggesting that purifying selection may minimize *trans*- divergence between species to a greater degree than *cis*- divergence (Wittkopp et al. 2008). Some aspects of regulatory evolution, however, may evolve under relaxed purifying selection, or even neutrally (Jordan et al. 2005). Moreover, expression of RAG1 and RAG2 may be under more severe purifying selection in tissues where it is highly expressed, such as bone marrow, than other tissues with lower levels of expression, such as heart. Overall, however, antagonistic *cis*- and *trans*- regulatory divergence, comparable levels of expression divergence, and similarities in the hierarchy of expression suggest the action of purifying selection on regulation of both of these genes in XL and XB.

Regulatory evolution could also be faster if purifying selection is relaxed after gene duplication. In the nematodes *Caenorhabdites elegans* and *C. briggsae*, for example, protein and regulatory divergence is weakly correlated in orthologs, but not in paralogs (Castillo-Davis et al. 2004). In *X. laevis*, divergence of paralogous expression profiles across multiple tissues and developmental stages is not correlated with molecular evolution of paralogous sequences, suggesting that selective constraints on protein sequence and expression are not coupled soon after WGD (Chain et al. 2008). WGD by allopolyploidization can lead to rapid or immediate changes in gene expression that can be nonadditive with respect to parental expression levels (Adams et al. 2003; Albertin et al. 2006; Wang et al.

2006). Additionally, rapid divergence of paralogous expression profiles on a quantitative, spatial, or temporal dimension, could contribute to their propensity to continue to encode functional proteins over the long term (Force et al. 1999; Lynch and Force 2000).

For these reasons, we expected that expression divergence between RAG1 orthologs would be greater than between RAG2 orthologs because the former but not the latter is an expressed duplicate (Prediction 4), but this expectation was not supported (Suppl. Table. 2). We found that the hierarchy of expression of RAG1 paralogs and RAG2β was qualitatively similar (Prediction 3, Table 5), which is consistent with these interacting proteins facing similar selective pressures. This result is echoed in fruit flies and yeast, where interacting proteins tend to have similar levels of intraspecific expression polymorphism and their expression levels are positively correlated across strains (Lemos et al. 2004). Similarities in the hierarchy of expression in RAG1 and RAG2 could be a consequence of co-evolution or co-regulation. The overall similarity in the direction of *trans*- divergence across genes, wherein the XB allele is generally upregulated (except in liver, Table 2) is consistent with (but not proof of) shared regulation in some tissue types. However, in liver significant divergence between orthologs was detected in parental individuals (RAG2β and in RAG1α) but in opposite directions in each gene: XL RAG2β is upregulated compared to XB RAG2β, possibly by *trans*- divergence whereas in liver *trans*- divergence upregulates XB RAG1α compared to XL RAG1α (Table 2). At least some aspects of *trans*- regulation of RAG1 and RAG2, therefore, are unique.

Conclusions

This study suggests that expression bias occurring soon after gene duplication has the potential to influence genomic restructuring (pseudogenization and gene deletion) that occurs afterwards. We speculate, therefore, that early changes in gene regulation could be particularly significant in sculpting the genetic fates of duplicate genes generated by WGD – even though these fates (pseudogenization, deletion, or persistent functionality) may be realized many millions of years later.

Additionally, we found that the magnitude and direction of orthologous and paralogous divergence can be tissue specific (Predictions 1 and 2). This suggests that variation in expression between different cell types within a single tissue – for example between epithelial and mesenchymal cell types in the liver that was analyzed in this study – could be modulated in unique and nonoverlapping ways. Inferences made about regulatory mechanisms from the entire body (Landry et al. 2005; Wittkopp et al. 2004) or from pooled differentiated cell types within an organ (this study; Cowles et al. 2002; Kirst et al. 2005) then, probably most strongly reflect a combined influence of regulatory mechanisms in (a) the cell type in which a gene is most highly expressed and (b) the cell type that constitutes the highest proportion of sample. Because of this, the inferred impact of these mechanisms actually might not correspond with those operating in any individual cell. This poses a challenge to the precise dissection of the mechanisms that drive expression divergence at the cellular level.

Acknowledgements

We thank Patricia Wittkopp for providing statistical advice and SAS scripts, Andrew Meade and Mark Pagel for advice on using BayesTraits, Adam Smith for assistance with pyrosequencing, Darcy Kelley and Elizabeth Leininger for providing some of the F1 hybrid individuals, and Mohammad Iqbal Setiadi and Frédéric Chain for assistance with animal propagation. This research was supported by the Canadian Foundation for Innovation, the National Science and Engineering Research Council, the Ontario Research and Development Challenge Fund, and McMaster University.

References

Adams KL, Cronn R, Percifeld R, Wendel JF (2003) Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. Proceedings of the National Academy of Sciences 100:4649

Albertin W, Balliau T, Brabant P, Chévre A-M, Eber F, Malosse C, Thiellement H (2006) Numerous and rapid nonstochastic modifications of gene products in newly synthesized *Brassica napus* allotetraploids. Genetics 173:1101-1113

Bassing CH, Swat W, Alt FW (2002) The mechanism and regulation of chromosomal V(D)J recombination. Cell 109:S45-S55

Brem RB, Yvert G, Clinton R, Kruglyak L (2002) Genetic dissection of transcriptional regulation in budding yeast. Science 296:752-755

Castillo-Davis CI, Hartl DL, Achaz G (2004) *cis*-regulatory and protein evolution in orthologous and duplicate genes. Genome Research 14:1530-1536

Chain FJJ, Evans BJ (2006) Molecular evolution of duplicate genes in *Xenopus laevis* is consistent with multiple mechanisms for their retained expression. PLoS Genetics 2:e56

Chain FJJ, Ilieva D, Evans BJ (2008) Duplicate gene evolution and expression in the wake of vertebrate allopolyploidization. BMC Evolutionary Biology 8:43

Cowles CR, Hirschhorn JN, Altshuler D, Lander ES (2002) Detection of regulatory variation in mouse genes. Nature Genetics 32:432-437

Denver DR, Morris K, Streelman JT, Kim SK, Lynch M, Thoman WK (2005) The transcriptional consequences of mutation and natural selection in *Caenorhabditis elegans*. . Nature Genetics 37:544-548

Evans BJ (2007) Ancestry influences the fate of duplicated genes millions of years after duplication in allopolyploid clawed frogs (*Xenopus*). Genetics 176:1119-1130

Evans BJ (2008) Genome evolution and speciation genetics of allopolyploid clawed frogs (*Xenopus* and *Silurana*). Frontiers in Bioscience:4687-4706

Evans BJ, Kelley DB, Melnick DJ, Cannatella DC (2005) Evolution of RAG-1 in polyploid clawed frogs. Molecular Biology and Evolution 22:1193-1207

Evans BJ, Kelley DB, Tinsley RC, Melnick DJ, Cannatella DC (2004) A mitochondrial DNA phylogeny of clawed frogs: phylogeography on sub-Saharan Africa and implications for polyploid evolution. Molecular Phylogenetics and Evolution 33:197-213

Felsenstein J (2005) PHYLIP (Phylogeny Inference Package) versin 3.6. Distributed by the author, Department of Genome Sciences, University of Washington, Seattle.

Force A, Lynch M, Pickett B, Amores A, Yan YL, Postlethwait JH (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531-1545

Greenhalgh P, Olesen CEM, Steiner LA (1993) Characterization and expression of Recombination Activating Genes (RAG-1 and RAG-2) in *Xenopus laevis.* Journal of Immunology 151:3100-3110

Huelsenbeck JP, Ronquist F (2001) MrBayes: Bayesian inference of phylogenetic trees. Bioinformatics 17:754-755

Jordan IK, Marino-Ramirez L, Koonin EV (2005) Evolutionary significance of gene expression divergence. Gene 345:119-126

Kirst M, Basten CJ, Myburg AA, Zeng ZB, Seberoff RR (2005) Genetic architecture of transcript-level variation in differentiating xylem of a eucalyptus hybrid. Genetics 169:2295-2303

Kruckeberg KE, Thibodeau SN (2004) Pyrosequencing Technology as a Method for the Diagnosis of Multiple Endocrine Neoplasia Type 2. Clinical Chemistry 50:522-529

Landry CR, Wittkopp PJ, Taubes CH, Ranz JM, Clark AG, Hartl DL (2005) Compensatory cis-trans evolution and the dysregulation of gene expression in interspecific hybrids of *Drosophila*. Genetics 171:1813-1822

Lemos B, Meiklejohn CD, Cáceres M, Hartl DL (2005) Rates of divergence in gene expression profiles of primates, mice, and flies: stabilizing selection and variability among functional categories. Evolution 59:126-137

Lemos B, Meiklejohn CD, Hartl DL (2004) Regulatory evolution across the protein interaction network. Nature Genetics 36:1059-1060

Ludwig MZ, Bergman C, Patel NH, Kreitman M (2000) Evidence for stabilizing selection in a eukaryotic enhancer element. Nature 403:564-567

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

Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG (2004) Genetic analysis of genome-wide variation in human gene expression. Nature 430:743-747

Nilsson TK, Johansson CA (2004) A novel method for diagnosis of adult hypolactasia by genotyping of the -13910 C/T polymorphism with Pyrosequencing technology. Scandanavian Journal of Gastroenterology 39:287-290

Organ CL, Shedlock AM, Meade A, Pagel M, Edwards SV (2007) Origin of avian genome size and structure in non-avian dinosaurs. Nature 446

Pagel M (1997) Inferring evolutionary processes from phylogenies. Zoologica Scripta 26:331-348

Pagel M (1999) Inferring the historical patterns of biological evolution. Nature 401:877-884

Rice WR (1989) Analyzing tables of statistical tests. Evolution 43:223-225

Spicuglia S, Franchini DM, Ferrier P (2006) Regulation of V(D)J recombination. Current Opinion in Immunology 18:158-163

Venditti C, Meade A, Pagel M (2006) Detecting the node-density artifact in phylogeny reconstruction. Systematic Biology 55:637-643

Wang J, Tian L, Lee H-S, Wei NE, Jiang H, Watson B, Madlung A, Osborn TC, Doerge RW, Comai L, Chen ZJ (2006) Genomewide nonadditive gene regulation in *Arabidopsis* allotetraploids. Genetics 172:507-517

Wittkoop PJ, Haerum BK, Clark AG (2004) Evolutionary changes in cis and trans gene regulation. Nature 430:85-88

Wittkopp PJ, Haerum BK, Clark AG (2004) Evolutionary changes in cis and trans gene regulation. Nature 430:85-88

Wittkopp PJ, Haerum BK, Clark AG (2008) Regulatory changes underlying expression differences within and between *Drosophila* species. Nature Genetics 40:346-350

Yan H, Yuan W, Velculescu VE, Vogelstein B, Kinzler KW (2002) Allelic variation in human gene expression. Science 297:1143

Yvert G, Brem RB, Whittle J, Akey JM, Foss E, Smith EN, Mackelprang R, Kruglyak L (2003) Trans-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. Nature Genetics 35:57-64

Figure Legends

Fig. 1. Expression divergence can occur between tissues, the sexes, duplicated genes (paralogs), and between species (orthologs), and is derived either from *cis*- and/or *trans*- acting factors. Intraspecific expression divergence between the sexes and between tissues must be caused by *trans*- acting factors because the *cis*- acting factors are identical within a species. An exception to this, indicated by one asterisk, is expression divergence between the sexes that arises due to differences in allelic copy number or the sex-determining locus. Because they share the same cellular environment, divergent expression of paralogs (α, β) that are co-expressed arises from *cis*- acting mechanisms. If paralogs of a gene are expressed in different species (i.e. not co-expressed), expression divergence between them could occur by *cis*-, *cis*- and *trans*-, or only *trans*- mechanisms. Divergence between orthologs in

different species (sp.1, sp.2) can arise by *cis*- or *trans*- acting factors, or both. Modified from (Evans 2008).

Fig. 2: Characterization of *cis*- and *trans*- contributions to expression divergence between species by comparing allelic expression ratios in parental and hybrid individuals. *Cis*- acting elements are depicted as 5' regulatory regions and trans-acting factors are depicted as transcription factors that bind to these elements. Inferences based on comparisons of expression ratios in parentals and hybrids assume intra-specific variation in expression is small relative to inter-specific variation. Only one allele is shown for the parental species but both are shown for hybrids. In this example, red elements up-regulate expression and blue colors down-regulate it. In hybrid individuals transcription factors from both parental species interact with *cis*- regulatory elements of the alleles from both parental species. In these comparisons, P refers to the $\log_2$-transformed expression ratio of parental alleles in each parental species and H refers to the $\log_2$-transformed expression ratio of each parental allele in a hybrid. *Cis*- divergence is indicated when the null hypothesis of P = H is not rejected but the null hypothesis of H = 0 is rejected. *Trans*- divergence is indicated when the null hypothesis of P = H is rejected but the null hypothesis of H = 0 is not rejected (which also implies rejection of the null hypothesis of P =0). However, if the parental expression ratio is closer to one than the hybrid expression ratio, this suggests that *cis*- and *trans*- acting factors have acted in opposite directions to stabilize expression levels in each species. When the expression ratio in hybrids is closer to one than the parental ratio, it may not be possible to distinguish between scenarios involving synergy and opposition of *cis*- and *trans*- acting factors.

Fig. 3. Putative allopolyploid evolution of the tetraploids *X. laevis* (XL), *X. borealis* (XB), and the tetraploid ancestors of *Xenopus* octoploids (Others). Daggers indicate extinct diploid ancestors or genes. Nodes 1 and 2 correspond with the divergence and union, respectively, of two diploid genomes; Node 3 indicates the MRCA of all *Xenopus* species and Node 4 refers to the MRCA of a subset of these species. (A) A reticulate phylogeny with ploidy in parentheses. (B) Nuclear genealogy assuming no recombination and no gene conversion between alleles at different paralogous loci (α and β). The dashed portion of the paralogous lineages in (B) evolved independently in different diploid ancestors. Octoploids, which have 72 chromosomes, inherited the complete genomes of two tetraploid ancestors, each of which had 36 chromosomes, so with no gene deletion they are expected to have two α and two β paralogs. Some species experienced gene deletion (Evans 2007) so the actual number of paralogs is less in these species. (C) Genetic distances between species that evolved through allopolyploidization can be estimated as the average of the branchlengths between each of the MRCAs. For instance, in this example phylogeny two an allotetraploid evolved from two extinct diploids. This allotetraploid then speciated into two descendant tetraploids (T1 and T2), ancestors of which underwent allopolyploidization to form an octoploid (O). The genetic distance between tetraploid T1 and the octoploid O is equal to the mean of the red and blue branchlengths. Each of these branches can be calculated from patristic distances from a genealogy that includes all paralogs in each species (see text for details).

Suppl. Fig. 1. Evolutionary relationships among clawed frogs reticulate but genetic distances can be used to represent these relationships with a bifurcating tree. (A) Inferred evolutionary relationships among 8 species based on analysis of RAG1 and RAG2 (Evans 2007). (B) Bifurcating phenogram based on genetic distances that was used in ancestral state reconstruction. In (B), numbered nodes correspond with those in Fig. 3A and 3B.

Table 1. Paralogous expression ratio in RAG1 in different species including tetraploids (4n) and octoploids (8n). Most log transformed expression ratios of RAG1α/RAG1β are positive, which is consistent with the SSRα hypothesis.

Species	Ploidy	Tissue	$\log_2$(expression ratio)
X. laevis	4n	bone marrow	0.1391
		liver	-0.0462
X. borealis	4n	bone marrow	-0.2108
		liver	-0.1992
X. gilli	4n	bone marrow	0.7985
X. muelleri	4n	bone marrow	0.9135
		heart	0.7117
		brain	3.2536
X. boumbaensis	8n	bone marrow	0.1519
X. andrei	8n	bone marrow	1.6814
X. itombwensis	8n	bone marrow	0.9425
		brain	-0.2954
		heart	0.3293
X. amieti	8n	bone marrow	0.9308
		heart	-0.1508
		liver	0.0051
		brain	-0.3201

Table 2. Log_2 transformed parental and hybrid expression ratios of (XL/XB) and probabilities of Type I error for rejecting the null hypotheses that the log transformed parental ratio is equal to zero (P = 0), that these ratios are equal (P = H), or that the log transformed hybrid ratio is equal to zero (H = 0). Interpretations include *trans-* divergence (T), and *cis-* divergence (C), and *cis-* and *trans-* divergence acting in opposite directions (CT antagonism). Inferences are based on two tests, (P = H) and (H = 0), and significant departures of each test after correction for two tests are indicated by asterisks. Individual significance of the test that the log transformed parental ratio is equal to zero (P < 0.05) is also indicated with an asterisk.

	Parental ratio	P value (P = 0)	Hybrid ratio	P value (P = H)	P value (H = 0)	Interpretation
RAG2β						
Bone Marrow	-0.3530	0.2931	0.8839	0.0061*	0.0002*	C: XL up, T: XB up
Brain	0.1330	0.6291	0.3861	0.4387	0.0441	almost significant C: XL up
Heart	0.1557	0.3688	1.1566	0.0012*	0.0001*	C: XL up, T: XB up
Liver	0.2264	0.0028*	0.0825	0.2570	0.4524	C and/or T: XL up(?)[a]
RAG1α						
Bone Marrow	-0.6271	0.1569	-0.2362	0.3643	0.2160	-
Brain	-0.8851	0.1728	-0.1330	0.2464	0.2941	-
Heart	-0.3148	0.0105*	0.2301	0.1087	0.4298	C and/or T: XB up(?)[a]
Liver	-0.2877	0.0032*	0.3522	0.0067*	0.0573	C:(?)[b]; T: XB up
RAG1β						
Bone Marrow	-0.2383	0.5413	0.2532	0.2563	0.0069*	C: XL up, T: XB up(?)[c]
Brain	-0.3846	0.3007	0.3058	0.1439	0.2435	-
Heart	-0.0537	0.7540	-0.2418	0.3967	0.1018	-
Liver	0.0229	0.7234	-0.1001	0.1889	0.1237	-

[a]Ambiguous interpretation because P = H and H = 0 not rejected

[b]Ambiguous interpretation because H = 0 not rejected

[c]Ambiguous interpretation because P = H not rejected

Table 3. Probability of Type I error for rejecting the null hypothesis of no difference in expression ratios of parental alleles in pairwise comparisons between tissue types in hybrids. For each gene, significantly different ratios after correction for six tests are indicated with an asterisk.

Comparisons between parental alleles (XL and XB) in hybrids			
RAG2β	Bone Marrow	Brain	Heart
Brain	0.0332		
Heart	0.1914	0.0019*	
Liver	0.0002*	0.1143	0.0001*
RAG1α			
Brain	0.6416		
Heart	0.1663	0.2599	
Liver	0.0248	0.0444	0.7018
RAG1β			
Brain	0.8323		
Heart	0.0017*	0.0640	
Liver	0.0001*	0.1252	0.2762

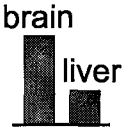
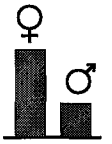
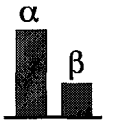
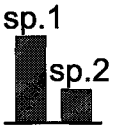
Table 4. Probability of Type I error for rejecting the null hypothesis of no difference in expression ratios of RAG1 paralogs in pairwise comparisons between hybrid tissue types. Significantly different ratios after correction for six tests are indicated with two asterisks and individually significant comparisons are indicated with one.

	XLα/XLβ		
hybrids	Bone Marrow	Brain	Heart
Brain	0.0363*		
Heart	0.0140*	0.2236	
Liver	0.0010**	0.1231	0.7356

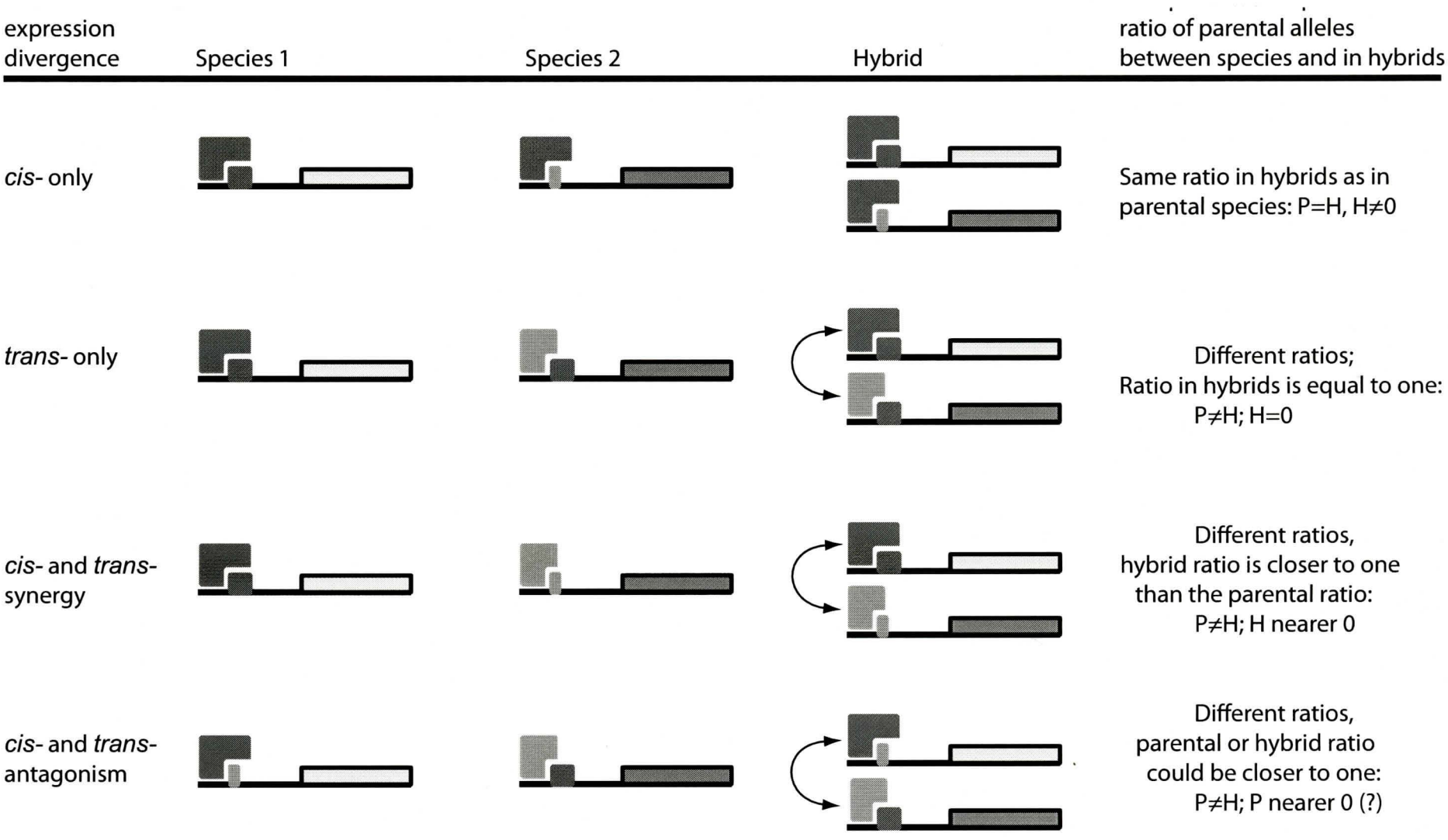
	XBα/XBβ		
hybrids	Bone Marrow	Brain	Heart
Brain	0.0144*		
Heart	0.2492	0.5407	
Liver	0.1399	0.0610	0.6745

Table 5: Fold difference in expression intensity within XL and XB in different tissue types. Asterisks indicate significant differences between tissue types. Some comparisons were not performed (NP) because of failed nucleotide extractions.

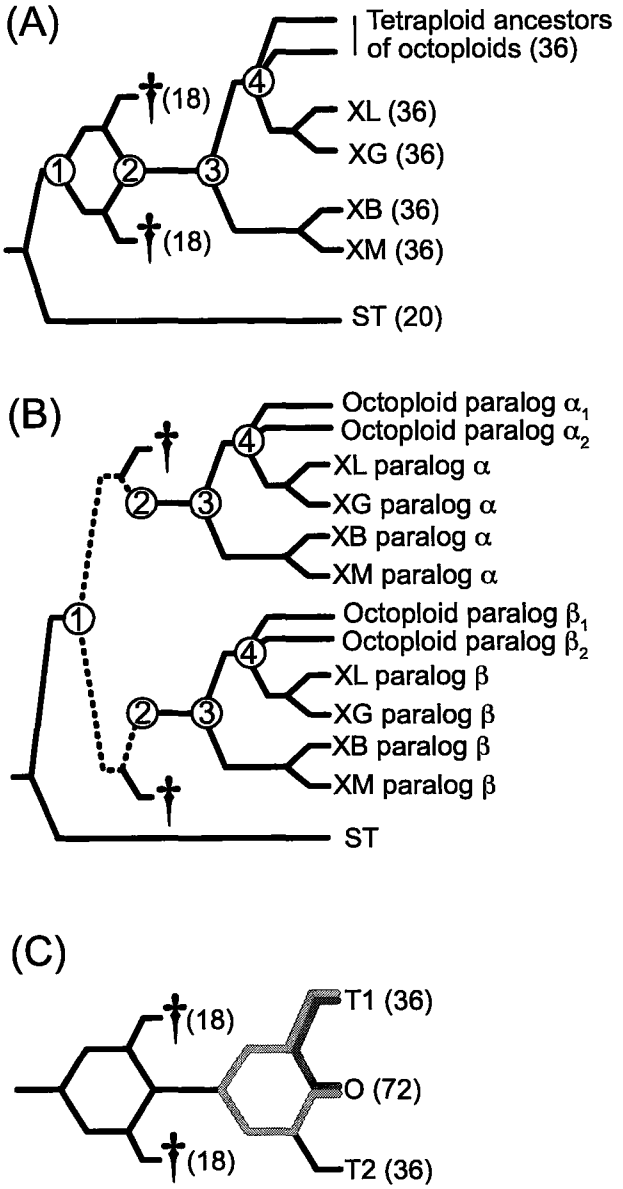
RAG2β	heart / liver	brain / heart	bone marrow / brain
XL	1.424	2.301*	1.005
XB	1.430	2.320*	1.530
RAG1α			
XL	1.460	1.537*	NP
XB	1.394	2.193*	NP
RAG1β			
XL	0.619	1.534	1.811
XB	0.725	1.285	1.129

Comparison		Mechanism for expression divergence
Between tissues		*trans-*
Between sexes		*trans-**
Between co-expressed paralogs		*cis-*
Between paralogs that are not co-expressed		*cis-* and/or *trans-*
Between orthologs		*cis-* and/or *trans-*

Anderson and Evans, Fig. 1



Anderson and Evans, Fig. 2



Anderson and Evans, Fig. 3

Supplementary Table 1: Expression ratios in parental and hybrids are not significantly different between the sexes, and we did not recover a significant parent of origin effect in hybrids. Indicated below are the probabilities of Type I error for rejecting the null hypothesis of no differences between comparisons. Parent of origin comparison was not performed on females because no data was obtained from female hybrids with an XB mother. Additionally, some comparisons were not possible (NP) because of failed nucleotide extractions.

Tests of individual genes				
	Parent of Origin effect test (in Hybrids)		Sex effect test	
RAG2	Male F1s only	Both sexes	Hybrids	Parents
Bone Marrow	0.1392	0.4972	0.8647	0.1992
Brain	0.6058	0.3770	0.6510	0.3148
Heart	NP	0.5432	0.9887	0.2607
Liver	0.3040	0.2717	0.8811	0.9768
RAG1 alpha				
Bone Marrow	NP	0.4354	0.4354	0.6170
Brain	0.3816	0.2663	0.8529	0.3662
Heart	NP	0.5484	0.5484	0.5586
Liver	0.4039	0.2384	0.3131	0.1498
RAG1 beta				
Bone Marrow	NP	0.0903	0.0903	0.9578
Brain	0.8436	0.7618	0.9652	0.8766
Heart	NP	0.7190	0.7190	0.7585
Liver	0.5353	0.2862	0.3262	0.1510

Tests of RAG1 paralogs				
	Parent of Origin effect test (in Hybrids)		Sex effect test	
	Male F1s only	Both sexes	Hybrids	Parents
XLα/XLβ				
Bone Marrow	NP	0.1750	0.1750	0.4882
Brain	0.7756	0.9775	0.6179	0.5286
Heart	NP	0.3895	0.3895	0.2198
Liver	0.8887	0.5503	0.2662	0.3846
XBα/XBβ				
Bone Marrow	NP	0.1873	0.1873	0.3160
Brain	0.3555	0.0773	0.4840	0.4226
Heart	NP	0.7965	0.7965	0.7744
Liver	0.3520	0.6620	0.7362	0.1319

Supplementary Table 2. Probability of Type I error for rejecting the null hypothesis that the absolute value of orthologous expression divergence of different genes are equal. Orthologous expression ratios used in these comparisons are listed in Table 1. Parental ratios (P) of RAG2β, RAG1α, and RAG1β are not significantly different from one another. Some non-independent pairwise comparisons between hybrid ratios (H), indicated by asterisks, are significantly different after correction for three tests per tissue type per treatment (P or H).

	BM		Brain		Heart		Liver	
Comparison	P	H	P	H	P	H	P	H
RAG2β vs. RAG1α	0.9839	0.0193*[a]	0.4840	0.0626	0.9627	0.0093*[a]	0.3766	0.0622
RAG2β vs. RAG1β	0.4947	0.0083*[a]	0.5203	0.7880	0.9440	0.0007*[a]	0.2423	0.3952
RAG1α vs. RAG1β	0.6404	0.7948	0.6533	0.3688	0.8837	0.9209	0.0482	0.0035*[b]

[a]RAG2β orthologs are more divergent

[b]RAG1α more divergent

Supplementary material 1. Pyrosequencing primers used to examine expression ratios of RAG1 and RAG2 orthologs and RAG1 paralogs.

RAG1

Forward Primer

5' GAGACCCTAACAGCCATTTTGAGTC 3'
Starts at position 2014

Reverse Primer (biotinylated)

5' CATACTGCTCTTCRACCACGCTCTY 3'
Ends at position 2136

Sequencing primer #1 (identifies XL alpha paralog)

5' CCATTTTGAGTCCTCTC 3'
Starts at position 2027
SNP target is at position 2049
C = XL alpha paralog
T = all other paralogs

Sequencing primer #2 (identifies XL beta and XB beta paralogs, and also distinguishes alpha and beta paralogs)

5' GGGAGGCTATGAAAAC 3'
Starts at position 2054
SNP target is at position 2073
C = XL beta paralog, *X. gilli* beta paralog, *X. andrei* beta 1 paralog, *X. amieti* beta 1 paralog, *X. boumbaensis* beta 1 paralog, *X. itombwensis* beta 1 and beta 2,
A = XB beta paralog, *X. muelleri* beta paralog
T = XL and XB alpha paralogs, *X. gilli* alpha paralog, *X. muelleri* alpha paralog, *X. andrei* alpha 1 and 2 paralogs, *X. amieti* alpha 1 and 2 paralogs, *X. boumbaensis* alpha 1 and 2 paralogs, *X. itombwensis* alpha 1 and alpha 2

Sequencing primer #3 (identifies XB alpha paralog)

5' GAAATKGGAGGAATTCT 3'
Starts at position 2086
SNP target is at position 2103
T = XB alpha paralog
A = all other paralogs

RAG2

Forward primer (biotinylated strand)

5' AAGGTGGGGAGATCACACAATA 3'
Starts at position 251

Reverse primer

5' TTATTTGGGGTCTTTCCTCCAT 3'
Ends at position 281

Sequencing primer

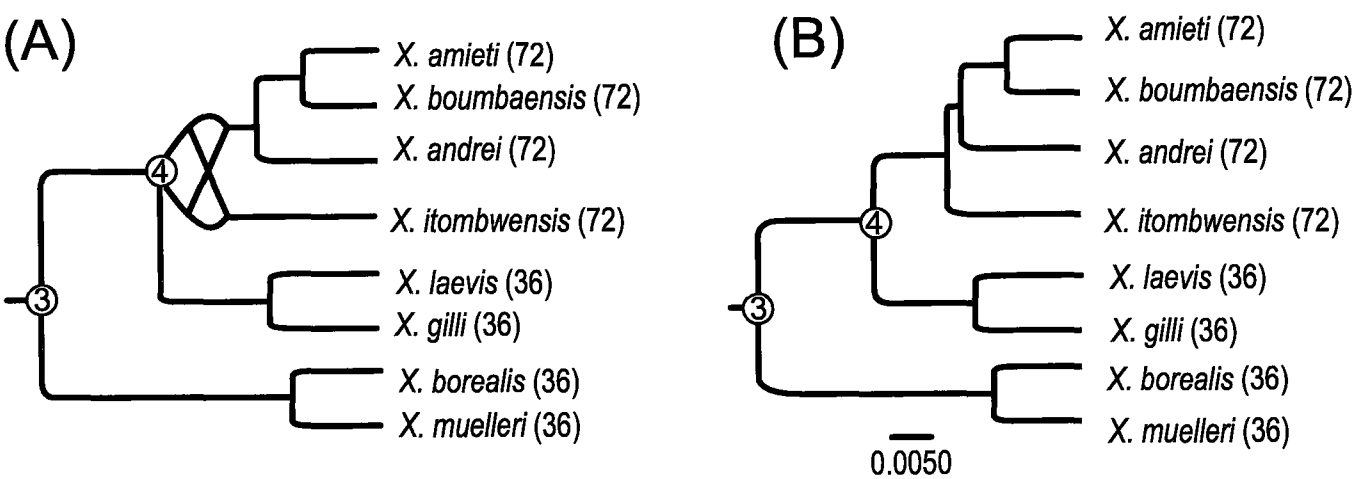
5' GGGTCTTTCCTCCATGT 3'

Ends at position 279

SNP target is at position 276

C = XL ortholog

G = XB ortholog



Anderson and Evans, Suppl. Fig. 1

Regulatory evolution of *dmrt1* in two developmental stages across species of allopolyploid clawed frogs (*Xenopus*).

Authors: Dave Anderson and Ben J. Evans[§]

Address: Center for Environmental Genomics, Department of Biology,
Life Sciences Building Room 328 McMaster University,
1280 Main Street West
Hamilton, ON, Canada L8S 4K1

[§]Corresponding author

Email addresses:

DA: andersdw@univmail.cis.mcmaster.ca

BJE: evansb@mcmaster.ca

Key Words: sex determination, *dmrt1*, cis- and *trans*- gene regulation, expression divergence, gene duplication, allopolyploidization, pseudogenization.

Abstract

The doublesex and mab-3 related transcription factor (DMRT1) plays a key role in sexual differentiation and/or sex determination in vertebrates. Using clawed frogs of the genus *Xenopus* as a model, we have examined mechanisms of expression divergence (i.e. *cis*- or *trans*-) of this crucial gene between species prior to and after primary sexual differentiation. In both developmental stages, expression divergence between species via *cis*- and *trans*- mechanisms is significant and antagonistic, but in opposite directions at each stage. These results suggest that DMRT1 expression at both of these developmental stages has been subject to purifying selection in the parental species and this expression is influenced by developmental-stage specific factors.

Introduction

Considerable variation exists among sexually reproducing species in the trigger for sexual differentiation, but the *doublesex- and mab-3 related transcription factor* (DMRT1) plays a remarkably conserved role in sexual differentiation among vertebrates. These genes contain a DNA-binding motif called a "DM domain" with sex specific function that have been identified in a variety of metazoans, including vertebrates, worms, flies, and coral (Miller et al. 2003; Yi and Zarkower 1999). In the medaka fish a DM domain- containing gene is also the primary sex-determining gene (Matsuda et al. 2007; Nanda et al. 2002), but in vertebrates generally it is downstream of other genetic or environmental mechanisms that trigger sexual differentiation (Haag and Doty 2005; Morrish and Sinclair 2002; Smith et al. 1999). In *C. elegans*, the DM gene *mab-3* regulates male-specific events including the suppression of yolk-synthesis and the development of male-specific cell lineages (Shen and Hodgkin 1988). In *D. melanogaster*, the DM domain gene *double sex* (*dsx*) is expressed in both male and female gonads, with one splice variant suppressing male differentiation and the other suppressing female differentiation (Burtis and Baker 1989). In the coral *Acropora millepora*, a gene with a DM domain, *AmDM1*, is upregulated during sex determination (Miller et al. 2003). In humans, it has been shown that homozygous deletions for *dmrt1* lead to gonadal sex-reversal, producing XY females (Veitia et al. 1998). That *dmrt1* is involved with sexual differentiation in amphibians is suggested by its upregulation in the primordial gonad and tissue-specific expression in adult testis of clawed frogs of the genus *Xenopus* (Osawa et al. 2005; Yoshimoto et al. 2006). Other male-specific or testis-specific genes that either suppress female differentiation (by inhibiting expression of the

"anti-testis" gene DAX1 (Yu et al. 1998) or promote male differentiation include Sry-like HMB box (SOX9), anti-Mullerian hormone (AMH), Wilms' tumor suppressor gene (WT1), and steroidogenic factor 1 (SF1) (Morrish and Sinclair 2002).

The pseudotetraploid clawed frog *Xenopus laevis* has two functional paralogous copies of *dmrt1*. Clawed frogs sexually differentiate via genetic mechanisms and females are the heterogametic (ZW) sex (Mikamo and Witschi 1963). Sexual differentiation in species with female heterogamy could be achieved either by a female-differentiation factor that is located on the W chromosome or by differences in dosage of a male determination or female repressor factor on the Z chromosome. In chickens, for example, DMRT1 is located on the Z-chromosome, females (which are heterogametic) have only one allele whereas males have two, and expression is higher in males (Nanda et al. 1999; Raymond et al. 1999). Recently, however, a new female-genome specific gene was reported that contains a DM domain (Yoshimoto et al. 2008). This gene is thought to be the primary trigger for female differentiation because transgenic animals that are genetically male (ZZ) but that carry the DM-W gene develop into females if this gene is expressed at an appreciable level (Yoshimoto et al. 2008). Because the N-terminal regions of the DM-W and DMRT1 genes, which contain a DM domain, are highly (89%) conserved, but the C-terminal is not homologous, one model of sex determination proposes that DM-W binds competitively with DMRT1 to inhibit transcription of genes involved with male sexual differentiation (Yoshimoto et al. 2008). In support of this model, DM-W is expressed in the primordial gonads just prior to DMRT1 in stage 50 female tadpoles (Yoshimoto et al. 2008). DMRT1 is expressed from tadpole stages 13-38,

and whole mount in situ hybridization indicates that expression of DMRT1 is restricted to the developing gonad by tadpole stage 52 – just prior to gonadal differentiation (Yoshimoto et al. 2006) but it is not expressed in the primordial gonads at an appreciable level until about tadpole stage 53 (Yoshimoto et al. 2008). In adults, DMRT1 is expressed only in testis although in juveniles it is also expressed at a lower level in ovary (Osawa et al. 2005).

Mechanisms of regulatory divergence of DMRT1

Because regulatory divergence of important genetic pathways could play a role in speciation, we have examined mechanisms of regulatory divergence between species of the DMRT1 gene. We studied how this gene is regulated in male tadpoles at a crucial stage of development – tadpole stage 53, which is soon after the primordial gonads are present but before they have undergone sexual differentiation. To test whether this gene is differently regulated through development, we compared these mechanisms to those involved with expression divergence in adult testes. More specifically, we compared the level of *cis*- divergence, which affects transcription in an allele specific manner, to the level of *trans*- divergence, which affects regulation of both alleles via direct or indirect interaction with *cis*- regulatory elements. We deployed an experimental design involving comparative expression analysis between two parental species and F1 interspecies hybrids (Anderson and Evans 2008; Landry et al. 2005; Wittkopp et al. 2004; Wittkopp et al. 2008).

Methods

Experimental approach

The relative contribution of *cis*- versus *trans*- acting mechanisms can be discerned through comparison of expression of species-specific alleles in two parental species and their F1 hybrids (Wittkopp et al. 2004). In hybrids, *trans*-acting factors from both species interact with the *cis*-acting factors of species-specific alleles, so their differential expression in these hybrids can be attributed exclusively to *cis*-divergence. In contrast, if divergent expression between species were only due to *trans*-acting mechanisms, both species-specific alleles would be expressed at the same level in hybrids, even though these alleles are expressed at different levels in each parental species. If *cis*- and *trans*- mechanisms both act to upregulate the expression of one species compared to another, the level of expression of species-specific alleles should be more similar in hybrids than in the parental species. Alternatively, if *cis*- and *trans*- acting factors drive expression divergence in opposite (antagonistic) directions, the expression level of species-specific alleles should be more similar in the parental species than in hybrids. Our experimental procedures, which follow (Landry et al. 2005; Wittkopp et al. 2004; Wittkopp et al. 2008) are described in detail elsewhere (Anderson and Evans 2008). In brief, we used a Biotage PSQ 96 pyrosequencer to quantify the ratio of expression of species-specific alleles in two species, *Xenopus laevis* and *Xenopus borealis* and also in their F1 hybrid.

We estimated the parental and hybrid expression ratio in the males at two developmental stages: (a) stage 53 tadpoles and (b) adult testis. Expression ratios were estimated in F1 hybrids generated from a cross between an XL female and an

XB male (6 tadpoles and 2 adults) or the reciprocal cross (4 tadpoles 2 adults), and compared to ratios the corresponding expression ratios in parental mixes (7 tadpole and 4 adult mixes). Primers and sequences are available in the supplemental information, and are based on previously acquired sequences from *X. laevis* (genbank accession numbers: *dmrt1*α - AB201112; *dmrt1*β - 252635) and our own cloned and sequenced samples for *X. borealis*. Because identification of males based on morphology was not possible, we used a genetic screen to identify male parental individuals and hybrids for the analysis of tadpole stage 53 (Supplemental information, figures 3.2 and 3.3).

Statistical analysis

We used the methods of (Landry et al. 2005) to test whether the parental and hybrid expression ratios were significantly different from each other, and to test whether expression ratios were significantly different from zero. We used restricted maximum likelihood (REML) to independently estimate variance parameters for each treatment being compared. Analyses were carried out using the "proc mixed" in SAS version 9.1.3 (SAS Institute) with modified scripts that were graciously provided by Patricia Wittkopp. Student's *t*-tests were computed within the mixed procedure and locus-level significance was interpreted after sequential Bonferroni correction for multiple tests (Rice 1989). Because an inference of *cis*- only divergence or *trans*-only divergence is made on the basis of rejection of one hypothesis but inability to reject another, these inferences suffer from Type II error (Anderson and Evans 2008). Neither inference, therefore, is statistically inconsistent with an inference of *cis*- and *trans*- mechanisms, which is based on rejection of two null hypotheses

(indicating that P ≠ H and H ≠ 0; Fig. 1). Application of the Bonferroni correction therefore makes the detection of antagonistic *cis*- and *trans*- divergence more conservative (Anderson and Evans 2008).

Results

Xenopus laevis and *X. borealis* both express two DMRT1 paralogs that originated by WGD.

We identified two paralogs of DMRT1 in *X. laevis* and in *X. borealis* and evolutionary relationships among them supports an origin by tetraploidization in *Xenopus* (i.e. clawed frogs with multiples of 18 chromosomes) (Fig. 1; Evans 2008). Analysis of sequence and expression of these genes indicates that both paralogs are functional and expressed in both of these species.

Mechanisms for expression divergence

We explored mechanisms of expression divergence by comparing expression ratios of species-specific alleles in parental and hybrid individuals. Tests for parent of origin effects on hybrid expression were not significant if corrected for multiple tests within both tadpole and adults for orthologous ratios of *dmrt1α*, *dmrt1β*, or for paralogous ratios of XL or of XB, and show significance without correction for multiple tests (P = 0.0435 for XL_{α}/XL_{β} and P = 0.0452 for *dmrt1β* in tadpoles, supplemental table 1). For subsequent comparisons we therefore report results from two separate analyses; the first analysis treats these data as biological replicates, and the second analysis separate tests for each type of hybrid cross (XL or XB mother).

In tadpole pre-differentiated male gonads, stage 53, significant upregulation of both XL paralogs of *dmrt1* is evident in parental species (P=0.0031 and P=0.0040 for α and β respectively; table 1). Both paralogs show evidence of trans upregulation of the XL copies (suggesting they may be regulated by common transcription factors, which are highly expressed in XL at this developmental stage), with the *trans*- test remaining significant even when treating the different hybrid crosses separately. Furthermore, the α paralog shows significant (or trending toward significant) antagonistic *cis*- divergence upregulating XB alleles, suggesting there may be compensatory expression divergence between these species, which would be consistent with (though cannot conclusively prove) the presence of either purifying or directional selection acting on expression during this stage. The β paralog approaches significance, but we cannot conclusively identify the same pattern of compensatory *cis*- divergence upregulating the XB copy.

In the adult testes, the α and β paralogs of *dmrt1* exhibit significantly different expression ratios in F1 hybrids (p=0.0001 and p=0.0079 respectively; table 3), with upregulation of the XL α paralog and upregulation of the XB β paralog. The *dmrt1α* paralog exhibits antagonistic *cis*- *trans*- divergence, with XL *cis*- upregulation and XB *trans*- upregulation (although the *cis*- component loses significance when the XB mothered hybrids are separated – this is most likely due to the small sample size of that group). It is noteworthy that this pattern of antagonistic divergence is exactly opposite of the pattern observed for this paralog in the tadpole stage, suggesting a largely different regulatory network during the adult stage. Reinforcing this is the

highly significant difference in paralog expression in the both tadpole and adult hybrids (supplemental table 2). The *dmrt1β* paralog shows evidence for the same pattern of antagonistic divergence as was seen for the α paralog, again suggesting a common set of transcription factors shared by both paralogs. However, the beta paralog divergence loses significance when the hybrid groups are separated.

Discussion and Conclusions

Natural selection can impact expression divergence in several ways. Purifying selection would cause overall expression levels remain constant (or with unexpectedly little variation) while different specific regulatory pathways may diverge between species (Ludwig et al. 2000). If expression divergence between species occurs in a neutral fashion, there should be a correlation between genetic distance and the magnitude of expression divergence. Directional selection acting on gene expression in two species could cause rapid divergence of regulatory mechanisms because one species may respond to the selection pressure with changes in *cis*-, while another might respond with changes in *trans*-. It is also possible to have an analogous resultant divergence via the same mechanism but via distinct changes, such that the differences are evinced in F1 hybrids. Moreover, parallel directional selection may be the fastest way for species to accumulate incompatible regulatory differences, such that there is dramatically lowered fitness upon hybridization (Johnson and Porter 2000).

Within a developmental stage, the pattern of expression divergence (e.g. *cis*- XL upregulation in adult testis) is consistent between both paralogs of DMRT1,

suggesting that many aspects of these regulatory networks (transcription factors and cis-regulatory modules) may be shared. Particularly interesting is the observation that there is actually directly opposite patterns of expression regulatory mechanism divergence in the tadpole stage as compared to the adult. This supports the notion that different regulatory networks may be active during different developmental stages, although they are still likely to contain many common elements (such as specific transcription factors, which may be expressed at different levels during the different stages).

The final conclusion to draw from this data is that the prevalence of antagonistic *cis*- and *trans*- acting divergence suggests that this gene has a high likelihood of hybrid misexpression, particularly during the physiologically crucial tadpole stage, when testis formation occurs. This makes it a promising candidate for physiological analysis in examining species divergence and reproductive isolation in the *Xenopus* lineage, and may play a significant role in speciation processes along other vertebrate lineages as well.

Implications for sex determination

Kobel (199^) found that unreduced eggs from a female F1 hybrid generated from a cross between *X. laevis* and *X. gilli* when backcrossed with sperm of *X. borealis* produced a clutch that was 98% male and with a putative sex chromosome genotype of $W_LZ_GZ_B$, where W and Z refer to the W and Z sex chromosomes and the subscript refers to the species from which it was derived (Kobel 1996). Likewise, this study also found that when eggs from this F1 female were backcrossed with *X. gilli* sperm to generate $W_LZ_GZ_G$ individuals, these individuals were mostly (88%) male. However,

when eggs from this F1 female were backcrossed with *X. laevis* or *X. muelleri* sperm to generate $W_LZ_GZ_L$ or $W_LZ_GZ_M$ individuals, all of these progeny were female (Kobel 1996). One explanation for this observation is that DMRT1 is upregulated in *X. borealis* and *X. gilli* to a level that is sufficient to counteract the feminizing effects of *X. laevis* DM-W when DMRT1 genes have twice the copy number that is normally in tetraploids. But our observations are inconsistent with this explanation, because we find that DMRT1 is downregulated in *X. borealis* compared to *X. laevis* at tadpole stage 53 (Table 1). It would be interesting, therefore, to better understand the mechanisms by which the protein products of these genes interact to orchestrate sexual differentiation, particularly in different species. One possibility, for example, is that the DM domain of *X. borealis* DMRT1 has particularly strong binding affinity to DNA, and that sex determination in each species is the evolutionary culmination of changes at the protein level of one gene that trigger or follow regulatory changes in another.

References

Anderson D, Evans BJ (2008) *Cis*- and *trans*- regulatory divergence between species, paralogs, and tissues: the RAG1/RAG2 heterodimer of polyploid clawed frogs. in preparation

Burtis KC, Baker BS (1989) *Drosophila doublesex* Gene Controls Somatic Sexual Differentiation by Producing Alternatively Spliced mRNAs Encoding Related Sex-Specific Polypeptides. Cell 56:997-1010

Evans BJ (2008) Genome evolution and speciation genetics of allopolyploid clawed frogs (*Xenopus* and *Silurana*). Frontiers in Bioscience:4687-4706

Haag ES, Doty AV (2005) Sex determination across evolution: connecting the dots. PLoS Biology 3:e21

Johnson NA, Porter AH (2000) Rapid speciation via parallel, directional selection on regulatory genetic pathways. Journal of Theoretical Biology 205:527-542

Kobel HR (1996) Allopolyploid speciation. In: Tinsley RC, Kobel HR (eds) The Biology of *Xenopus*. Clarendon Press, Oxford, p 391-401

Landry CR, Wittkopp PJ, Taubes CH, Ranz JM, Clark AG, Hartl DL (2005) Compensatory cis-trans evolution and the dysregulation of gene expression in interspecific hybrids of *Drosophila*. Genetics 171:1813-1822

Ludwig MZ, Bergman C, Patel NH, Kreitman M (2000) Evidence for stabilizing selection in a eukaryotic enhancer element. Nature 403:564-567

Matsuda M, Shinomiya A, Kinoshita M, Suzuki A, Kobayashi T, Paul-Prasanth B, Lau E-I, Hamaguchi S, Sakaizumi M, Nagahama Y (2007) DMY gene induces male development in genetically female (XX) medaka fish. Proceedings of the National Academy of Sciences 104:3865-3870

Mikamo K, Witschi E (1963) Functional sex-reversal in genetic females of *Xenopus laevis*, induced by implanted testes. Genetics 48:1411-1421

Miller SW, Hayward DC, Bunch TA, Miller DJ, Ball EE, Bardwell VJ, Zarkower D, Brower DL (2003) A DM domain protein from a coral, *Acropora millepora*, homologous to proteins important for sex determination. Evolution and Development 5:251-258

Morrish BC, Sinclair AH (2002) Vertebrate sex determination: many means to an end. Reproduction 124:447-457

Nanda I, Kondo M, Hornung U, Asakawa S, Winkler C, Shimizu A, Shan Z, Haaf T, Shimizu N, Shima A, Schmid M, Schartl M (2002) A duplicated copy of DMRT1 in the sex-determining region of the Y chromosome of the medaka *Oryzias latipes*. Proceedings of the National Academy of Sciences 99:11778-11783

Nanda I, Shan Z, Schartl M, Burt DW, Koehler M, Nothwang H, Grutzner F, Raton IR, Windsor D, Dunn I, Engel W, Staeheli P, Mizuno S, Haaf T, Schmid M (1999) 300 million years of conserved synteny between chicken Z and human chromosome 9. Nature Genetics 21:258-259

Osawa N, Oshima Y, Nakamura M (2005) Molecular cloning of *Dmrt1* and its expression in the gonad of *Xenopus*. Zoolog. Sci. 22:681-687

Raymond CS, Kettlewell JR, Hirsch B, Bardwell VJ, Zarkower D (1999) Expression of *Dmrt1* in the genital ridge of mouse and chicken embryos suggests a role in vertebrate sexual development. Developmental Biology 215:208-220

Rice WR (1989) Analyzing tables of statistical tests. Evolution 43:223-225

Shen MM, Hodgkin J (1988) *mab-3*, a Gene Required for Sex-Specific Yolk Protein Expression and a Male-SPecific Lineage in *C. elegans*. Cell 54:1019-1031

Smith CA, McClive PJ, Western PS, Reed KJ, Sinclair AH (1999) Conservation of a sex-determining gene. Nature 402:601-602

Veitia RA, Nunes M, Quintana-Murci L, Rappaport R, Thibaud E, Jaubert F, Fellous M, McElreavey K, Goncalves J, Silva M, Rodrigues JC, Caspurro M, Boieiro F,

Marques R, Lavinha J (1998) Molecular Analysis of 9p Deletions Associated with XY Sex Reversal: Refining the Localization of a Sex-Determining Gene to the Tip of the Chromosome. American Journal of Human Genetics 63:905-908

Wittkopp PJ, Haerum BK, Clark AG (2004) Evolutionary changes in cis and trans gene regulation. Nature 430:85-88

Wittkopp PJ, Haerum BK, Clark AG (2008) Regulatory changes underlying expression differences within and between *Drosophila* species. Nature Genetics 40:346-350

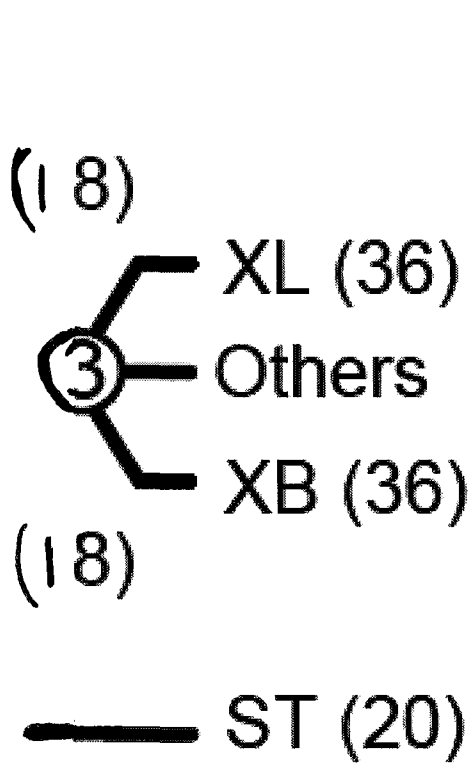
Yi W, Zarkower D (1999) Similarity of DNA binding and transcriptional regulation by *Caenorhabditis elegans* MAD-3 and *Drosophila melanogaster* DSX suggests conservation of sex determining mechanisms. Development 126:873-881

Yoshimoto S, Okada E, Oishi T, Numagami R, Umemoto H, Tamura K, Kanda H, Shiba T, Takamatsu N, Ito M (2006) Expression and promoter analysis of *Xenopus DMRT1* and functional characterization of the transactivating property of its protein. Develpment, Growth and Differentiation 48:597-603

Yoshimoto S, Okada E, Umemoto H, Tamura K, Uno Y, Nishida-Umehara C, Matsuda Y, Takamatsu N, Shiba T, Ito M (2008) A W-linked DM-domain gene, DM-W, participates in primary ovary development in *Xenopus laevis*. Proceedings of the National Academy of Science 105:2469-2474

Yu RN, Ito M, Saunders TL, Camper SA, Jameson JL (1998) Role of *Ahch* in gonadal development and gametogenesis. Nature Genetics 20:353-357

Figure 1. Evolutionary relationships among DMRT1 paralogs of *X. laevis* and *X. borealis* support an origin by tetraploidization in *Xenopus*. A phylogeny and genealogy of DMRT1 illustrates relationships among paralogs α and β of *X. laevis* (XLα and XLβ) and of *X. borealis* (XBα and XBβ) and an ortholog of *S. tropicalis* (ST). Dotted lines represent diploid genes, red lines are tetraploid genes immediately after duplication, and blue lines are tetraploid genes at a later stage after duplication. Labeled nodes refer respectively to (0) divergence of Silurana and Xenopus, (1) speciation of 2n=18 diploids in Xenopus, (2) allotetraploidization in Xenopus to create a 4n=36 ancestral species, (3) speciation of this species into allotetraploid descendants such as *X. laevis* and *X. borealis*. Daggers indicate diploid species that are presumed extinct.


(18)
XL (36)
3
Others
XB (36)
(18)
ST (20)

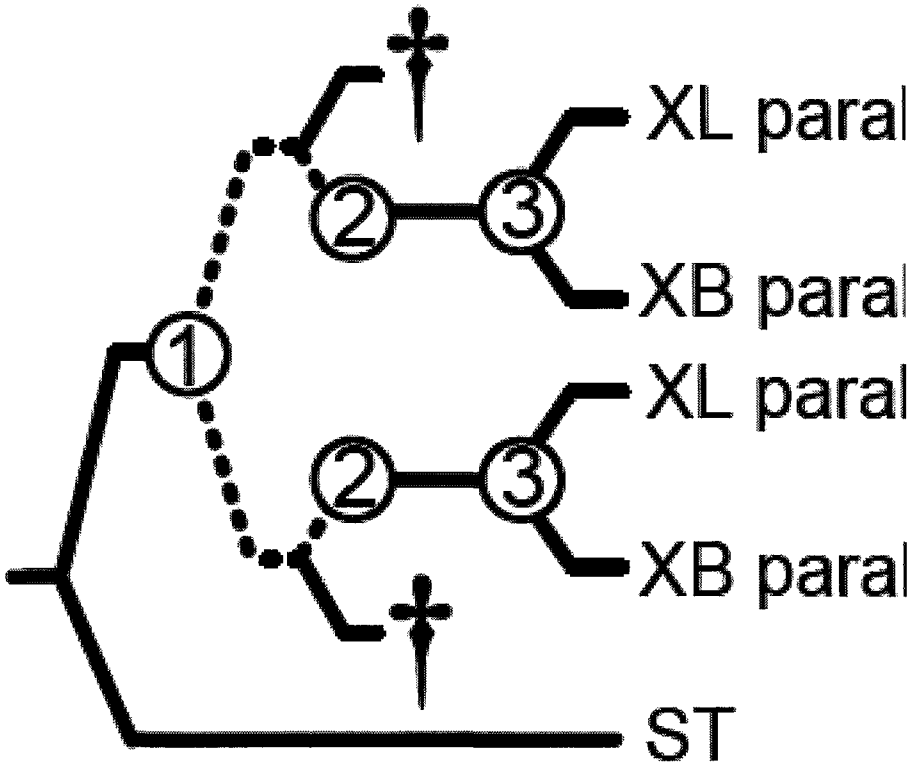

XL paral
2
3
XB paral
1
XL paral
2
3
XB paral
ST

Table 1. Log transformed parental and hybrid expression ratios of (XL/XB) and probabilities of the null hypothesis that the log transformed parental ratio is equal to zero (P = 0), that these ratios are equal (P = H), and that the log transformed hybrid ratio is equal to zero (H = 0). Interpretations include *trans*- divergence (T), and *cis*- divergence (C), and *cis*- and *trans*- divergence acting in opposite directions (CT antagonism). Inferences are based on two tests, (P = H) and (H = 0), and significant departures of each test after correction for two tests are indicated by asterisks. Individual significance of the test that the log transformed parental ratio is equal to zero (P < 0.05) is also indicated with an asterisk.

	Parental ratio	P value (P = 0)	Hybrid ratio	P value (P = H)	P value (H = 0)	Interpretation
DMRTα						
Stage 53, Tadpole	0.8388	0.0031*	-0.4173	0.0003*	0.0039*	C: XB up, T: XL up
Adult, Teste	0.6503	0.2864	2.9950	0.0097*	0.0001*	C:XL up, T: XB up
DMRTβ						
Stage 53, Tadpole	0.8570	0.0040*	0.1415	0.0283*	0.3576	T: XL up
Adult, Teste	-0.1452	0.7032	1.5794	0.0370*	0.0220*	C: XL up, T: XB up
[a]Ambiguous interpretation because P = H not rejected						
For Hybrids with XL mother						
DMRTα						
Stage 53, Tadpole	0.8388	0.0031*	-0.5797	0.0005*	0.0060*	C: XB up, T: XL up
Adult, Teste	0.6503	0.2864	3.4686	0.0072*	0.0001*	C:XL up, T: XB up
DMRTβ						
Stage 53, Tadpole	0.8570	0.0040*	0.3477	0.1253	0.0634	T: XL up
Adult, Teste	-0.1452	0.7032	1.7868	0.0734	0.0819	-
For Hybrids with XB mother						
DMRTα						
Stage 53, Tadpole	0.8388	0.0031*	-0.1738	0.0030*	0.0585	C: XB up (?), T: XL up
Adult, Teste	0.6503	0.2864	2.5214	0.0368*	0.0005*	C:XL up, T: XB up
DMRTβ						
Stage 53, Tadpole	0.8570	0.0040*	-0.2710	0.0032*	0.0960	T: XL up
Adult, Teste	-0.1452	0.7032	1.1644	0.1741	0.1561	-

Table 2. Probability of Type I error for rejecting the null hypothesis of no difference in expression ratios of parental alleles in pairwise comparisons between tissue types in parents and in hybrids. For each gene, significantly different ratios after correction for two tests are indicated with an asterisk.

Comparisons between parental alleles (XL and XB) in parents	
DMRT1α	Adult Teste
Tadpole, stage 53	0.7622
DMRT1β	
Tadpole, stage 53	0.0462

Comparisons between parental alleles (XL and XB) in hybrids	
DMRT1α	Adult Teste
Tadpole, stage 53	0.0001*
DMRT1β	
Tadpole, stage 53	0.0079*

Supplementary Table 1: Expression ratios in parental and hybrids are not significantly different between the sexes, and we did not recover a significant parent of origin effect in hybrids. Indicated below are the probabilities of Type I error for rejecting the null hypothesis of no differences between comparisons. Parent of origin comparison was not performed on females because no data was obtained from female hybrids with an XB mother. Additionally, some comparisons were not possible (NP) because of failed nucleotide extractions.

Tests of individual genes	
	Parent of Origin effect test (in Hybrids)
DMRT1 alpha	Male F1s only
Tadpole, stage 53	0.1056
Adult Testes	0.0606
DMRT1 beta	
Tadpole, stage 53	0.0452
Adult Testes	0.6826

Tests of DMRT1 paralogs	
	Parent of Origin effect test (in Male F1s only
XLα/XLβ	
Tadpole, stage 53	0.0435
Adult Testes	0.2899
XBα/XBβ	
Tadpole, stage 53	0.4020
Adult Testes	0.7276

Supplementary Table 2. Probability of Type I error for rejecting the null hypothesis that the absolute value of orthologous expression divergence of different genes are equal. Orthologous expression ratios used in these comparisons are listed in Table 1. Parental ratios (P) of RAG2β, RAG1α, and RAG1β are not significantly different from one another. Some non-independent pairwise comparisons between hybrid ratios (H), indicated by asterisks, are significantly different after correction for three tests per tissue type per treatment (P or H).

	Tadpole, stage 53		Adult Testes	
Comparison	P	H	P	H
DMRT1α vs. DMRT1β	0.9595	0.0062*	0.2554	0.0220*

Primer Sequences

Supplementary material 1. Pyrosequencing primers used to examine expression ratios of DMRT1.

DMRT1

Forward Primer

5' GGAATMAGCCATCCWATCCMTTTGC 3'

Starts at position 4693

Reverse Primer (biotinylated)

5' TGTKGAACCTGAAGTGGGTGTGC 3'

Ends at position 4805

Sequencing primer #1 (identifies XB beta paralog)

5' GGAATMAGCCATCCWATCCMTTTGC 3'

Starts at position 4693

SNP target is at position 4719

T = XB beta paralog

C = all other paralogs

Sequencing primer #2 (identifies XB alpha paralog)

5' ATCCWATCCMTTTGCCYATTGCAGC 3'

Starts at position 4703

SNP target is at position 4728

C = XB alpha paralog

T = All other paralogs

Sequencing primer #3 (identifies XL alpha paralog)

5' TCTTGCTTGATGYTGGAAARCAGT 3'

Starts at position 4765

SNP target is at position 4791

T = XL alpha paralog

C = all other paralogs

Sequencing primer #4 (identifies XB alpha paralog)

5' TGATAAAAARGGARCATGGTGGTAGCA 3'

Starts at position 4736

SNP target is at position 4764

T = XB alpha paralog

C = all other paralogs

II:

Concluding Statements

This thesis contributes to our understanding of gene expression regulatory mechanisms by examining expression divergence in two distinct systems – the *rag1-rag2* heterodimer and the *dmrt1* gene.

The study focusing on the genes *rag1* and *rag2* identifies tissue-specific divergence of expression regulatory mechanisms, suggesting that information may be obscured or lost by conducting whole-body assays, and the expression patterns detected likely reflect those of the predominant tissue in which the gene of interest is expressed. The study focusing on *dmrt1* (Chapter 3) suggests a potential role influencing hybrid male sterility in *Xenopus* clawed frogs. Given that this gene has demonstrated functional importance in many vertebrate species, including humans, the pattern of change in its regulatory network (and potential for directional sexual selection acting on it) has implications for many diverse sexually differentiating species. The observed expression divergence (and often resultant hybrid misexpression) has implications for the study of speciation, particularly via post-zygotic reproductive isolation.

It is clear that comparative analysis of the sort employed for this thesis allows one to elucidate the divergence of different mechanisms of gene expression in a gene specific manner. PSQ pyrosequencing offers a useful method to look at a gene of particular interest, and to examine in depth the quantitative detail of the manner in which it is expressed, particularly from an evolutionary standpoint.

There is an important contribution to be made by this method to overall evolutionary theory, as it explores in depth the relatively poorly understood connection between gene sequence and phenotype.

These methods also have applications to medical science, as there are several diseases for which the cause is known to be abnormal levels of gene expression, such as breast cancer (van 't Veer et al. 2002) and retinal diseases (Blackshaw et al. 2001). By conducting this type of analysis on these misexpressed genes, one uncovers details about the specific mechanisms that regulate their expression, and therefore one may better understand the distal cause of the disease. Such an understanding could be vital to designing expression regulatory treatments, such as those proposed for some types of schizophrenia (Mimics et al. 2001).

A large scale genome wide expression study that employs 454 pyrosequencing, as outlined by (Torres et al. 2008), could greatly expand the scope of this study. Provided the model system (i.e. organisms) have sufficient sequence divergence (and a sufficiently detailed EST database), it is possible to use this system to identify the relative proportions of species specific gene alleles, and thus tease apart the relative contributions of *cis*- and *trans*- acting regulatory factors to their overall expression differences.

References

Blackshaw S, Fraioli RE, Furukawa T, Cepko CL (2001) Comprehensive Analysis of Photoreceptor Gene Expression and the Identification of Candidate Retinal Disease Genes. Cell 107:579-589

Mimics K, Middleton FA, Lewis DA, Levitt P (2001) Analysis of complex brain disorders with gene expression microarrays: schizophrenia as a disease of the synapse. Trends in Neuroscience 24:479-486

Torres TT, Metta M, Ottenwalder B, Schlotterer C (2008) Gene expression profiling by massively parallel sequencing. Genome Research 18:172-177

van 't Veer LJ, Dai H, Vijver MJvd, He YD, Hart AAM, Mao M, Peterse HL, Kooy Kvd, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415:530-536

Wittkopp PJ, Haerum BK, Clark AG (2008) Regulatory changes underlying expression differences within and between *Drosophila* species. Nature Genetics 40:346-350

Appendix:

Pyrosequencing Methods

Methods for PSQ 96 Pyrosequencer

These methods are customized for use with 27 uL volume of PCR amplified product.

Materials

-Streptavidin Sepharose beads

-Vacuum prep tool

-70% Ethanol

-Biotage 40-0033 Binding Buffer

-Biotage 40-0034 Annealing Buffer

-Biotage 40-0035 Denaturation Solution

-Biotage 40-0036 Washing Buffer

-Sequencing Primer (at 20 μM concentration)

Methods

1. Before beginning prep, ensure that run has been entered into software program, and appropriate wells are activated.

2. Add Streptavidin Sepharose beads, Binding Buffer and ddH_20 to each PCR sample in these proportions:

   -3 uL sepharose beads (be sure to shake bottle thoroughly prior to adding beads)

   -13 uL ddH_20

   -37 uL binding buffer

3. Using a mixer/shaker, incubate at room temperature for at least 5 minutes agitating constantly to keep beads from settling (do not allow samples to sit unagitated for more than 3 minutes before proceeding to the step 8).

4. While samples are shaking, prepare PSQ 96 low-plate with 39 uL Annealing buffer and 1 uL sequencing primer per sample well.

5. Fill the vacuum prep troughs with 180 mL 70% ethanol, 120 mL denaturation solution, 180 mLwashing buffer and and 180 mL ddH_20.

6. Apply vacuum to the Vacuum Prep Tool.

7. Wash the probes of the Vacuum Prep Tool by lowering the tool into ddH_20 (parking position) for approximately 20-30 seconds. Make sure that proper vacuum has been attained.

8. Capture the beads containing immobilized templates on the filter probes by slowly lowering the Vacuum Prep Tool into the PCR plate (nt! – must take place within 3 minutes of terminating the agitation, as sepharose beads will quickly settle).

9. Once the liquid has been aspirated evenly from all wells, and all beads captured on probe tips, move the Vacuum Prep Tool to the trough containing 70% ethanol and let the solution flush through the filters for 5 seconds.

10. Move the Vacuum Prep Tool to Denaturation solution and flush through filters for 5 seconds.

11. Move the Vacuum Prep Tool to Washing buffer and flush through the filters for 5 seconds.

12. Allow liquid to completely drain from the probes, pick up the tool and move to beyond 90° vertical, and hold for 30 seconds before returning to horizontal position.

13. Turn off vacuum, and detach Vacuum Prep Tool (to ensure there is no more suction through the probe filters)

14. Release the beads in the pre-loaded PSQ 96 low-plate. Set Vacuum prep tool aside.

15. Heat the PSQ 96 low-plate (containing samples) to 90°C for 5 minutes.

16. Remove the plate from the heating block and let the samples cool to room temperature (approximately 5 minutes).

17. Load PSQ 96 cartridge with appropriate amounts of nucleotides, substrate and enzyme (calculated, and displayed with parameter details by PSQ software)

18. Place the cartridge and PSQ 96 low-plate with the samples in the PSQ 96 pyrosequencer.

19. Start the Run.