Gene dosage and the evolution of gene expression

Fabian Zimmer

Department of Genetics, Evolution and Environment

University College London (UCL)

A thesis submitted for the degree of

Doctor of Philosophy

I, Fabian Zimmer, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Fabian Zimmer

Abstract

The duplication and loss of genes, chromosomes and whole genomes has had a major impact on the evolution of most organisms. Changes in gene copy number, called gene dosage, may influence the resulting level of gene product through changes in gene expression. These gene expression changes can be detrimental, resulting in compensation and buffering mechanisms, or beneficial, when selection favours increased gene dosage. Understanding how changes in gene dose can influence the evolution of gene expression within and between species is an important task in evolutionary biology. This thesis combines studies of gene, protein domain, and genome duplications with gene expression data from a range of bird species to understand the evolutionary consequences of gene dosage changes.

In addition to gene duplication and loss events, the genomic location of genes can subject loci to different evolutionary pressures. Genes present on sex chromosomes or the mitochondria are inherited unequally between males and females, potentially causing sexual conflict over expression. This thesis investigates if inter-genomic conflict could drive gene movement on and off the sex chromosomes using a comparative genomics approach.

Acknowledgments

First and foremost, I would like to thank my supervisor ***Professor Judith Mank*** for her great support over the last few years. Her enthusiasm for biology has motivated me countless times. ***Professor Kevin Fowler*** has always offered invaluable advice and I could not imagine a better secondary supervisor. I would also like to commend ***Dr Peter Harrison*** for his help with computational problems and his feedback on all my manuscripts. My collaboration with ***Dr Rebecca Dean*** was a great working experience and I would like to thank her for supporting me with all my projects. I am indebted to ***Dr Stephen Montgomery*** who has gone above and beyond in helping me to resolve many scientific issues. Thanks to him, I know more about brains (including my own), phylogenies and cephalopods than I ever expected. I would like to thank all other members of the ***Mank lab*** (***Alison***, ***Vicencio, Natasha, Jen, Marie, Iulia, Clara***) and ***Professor Christophe Dessimoz*** for the many productive discussions – I enjoyed working with all of you. I would not have finished this thesis without the support of my ***friends and family***. In many regards writing a thesis is like cycling in the mountains and I would especially like to thank my friend ***Salvador Estrugo***, who has taught me how to climb (slowly). Last, but by no means least, I am infinitely grateful for all the support from my wife ***Tammela Platt*** – you helped me in more ways than I can list here.

To my parents

Table of Contents

Chapter 1 - Introduction 19

Gene dosage.. 21

Genome evolution 22

Whole genome duplications 22
Evolution after WGD events 24
Detection of WGD events.. 27

Gene dosage and sex chromosome evolution 29

Paralog divergence 32

Protein domain evolution 36

Gene movement and rearrangements 37

Summary of aims 39

Summary of thesis chapters 40

Glossary 41

Abbreviations 42

Additional published work 43

Chapter 2 - Compensation of dosage-sensitve genes on the chicken Z chromosome 45

Summary.. 48

Introduction.. 49

Material and Methods 52

RNA-Seq analysis and gene expression estimates 52
Identification of ohnologs and other paralogs 53
Functional annotation of ohnologs... 54
SNP calling and estimation of allele specific expression.............................. 54

Results 56

Incomplete dosage compensation in females and reduced Z expression in males 56
Z:A ratio comparison... 61
Expression level and dosage compensation 61
Allele-specific expression and potential for Z chromosome inactivation 64
Ohnologs are preferentially dosage-compensated....................................... 66
Older Z chromosome parts contain fewer ohnologs 72
Dosage compensation of ohnologs across tissues....................................... 73
Functional annotation of Z-linked ohnologs 75

Discussion.. 77

Chapter 3 - Inferring paralogs and expression divergence across multiple bird species using RNA-Seq data 79

Summary 81

Introduction 82

Material and Methods 86

RNA-Seq assembly and gene expression calculation 88
RNA-Seq isoform filtering and protein prediction 90
DNA-Seq data processing 92
Reconstruction of gene family evolution 92
Lineage-specific paralog sequence divergence 93

Results 94

Are inferred evolutionary events from RNA-Seq and DNA-Seq comparable? 94
Can RNA- and DNA-Seq data be combined to improve paralog detection? . 95
Can lineage-specific paralogs be filtered using ancestral family structure? .. 101
Is the duplication rate of inferred lineage-specific paralogs comparable to genomic estimates? 104
Can strictly filtered paralogs be confirmed using Ensembl data? 105

Discussion 107

Family structure filtering and duplication rate comparisons 108
Paralog pairs show low sequence identity and unexpected gene tree topology 109
An explanation: variation in gene model prediction 110

Chapter 4 - The potential role of sexual selection and sexual conflict on the genomic distribution of mito-nuclear genes 113

Summary 116

Introduction 117

Material and Methods 120

Detection and localization of genes interacting with mitochondria 120
Statistical analysis 122

Results and Discussion 124

Chapter 5 - Deficit of mito-nuclear genes on the human X chromosome predates sex chromosome formation 135

Summary 138

Introduction 139

Materials and Methods 142

Identification of ancestral chromosomes to the human X chromosome 142
Identification of mt-N gene movement 142
mt-N abundance 143

Results and Discussion 144

Mt-N gene movement on and off the human X chromosome 144
mt-N gene abundance on X syntenic regions 146
Gene annotation and mt-N abundance 148
Mt-N gene abundance across independent X chromosomes 149

Chapter 6 - Phylogenetic analysis supports a link between DUF1220 domain number and primate brain expansion 151

Summary 154

Introduction 155

Materials and Methods 158

Counting DUF1220 domains 158
Phylogenetic gene-phenotype analysis 158

Results 160

Variation in DUF1220 domain content 160
Exploring variation in evolutionary rate of DUF1220 domain number 162
Gene-Phenotype co-evolution 164
DUF1220 analysis using a Bayesian approach 165
DUF1220 analysis using Phylogenetic Generalised Least Square (square-root transformed counts) 170
DUF1220 analysis using Phylogenetic Generalised Least Square (*log*-transformed DUF1220 domain data) 172

Discussion 173

Chapter 7 - Discussion 177

Dosage compensation and whole genome duplications 180

Inferring gene family evolution from RNA-Seq data 181

Gene movement and mitochondrial interactions 184

DUF1220 domain dose increase 185

Analysing comparative transcriptome data: general lessons 188

Annotation and assembly quality differences 189

General limitations in bioinformatics analyses 191

References 193

Appendix 225

Table of Figures

Figure 2.1 Comparison of gene expression for auto. and Z-linked genes.......59

Figure 2.2 Expression level across all chromosomes for males and females..60

Figure 2.3 Z:A ratio across CPM thresholds for all males and all females.......62

Figure 2.4 *Log*$_2$ transformed CPM expression for genes on the Z chr.............63

Figure 2.5 Comparison of *log*$_2$-transformed fold-change between female and male expression for ohnologs and non-ohnologs on the Z chromosome.......68

Figure 2.6 Comparison of *log*$_2$-transformed fold-change between female and male expression for ohnologs taken from the strict OhnologsDB....................71

Figure 2.7 Dosage compensation overlap between ohnologs.........................74

Figure 3.1 Design of the bioinformatics pipeline...87

Figure 3.2 Cladogram of HOG family reconstruction RNA-Seq DNA-Seq.......96

Figure 3.3 Distribution of predicted gain, loss and duplication 97

Figure 3.4 Distribution of branch lengths..98

Figure 3.5 Cladogram of all used RNA-Seq and DNA-Seq assemblies...........99

Figure 3.6 Histogram of the % protein sequence identity between paralogs 103

Figure 4.1 Bias of nuclear-mitochondrial genes on the sex chromosomes...125

Figure 4.2 Bias of nuclear-mitochondrial genes on the sex chromosomes...126

Figure 4.3 Bias of nuclear-mitochondrial genes on the sex chromosomes across *G. aculeatus* and *O. anatinus*..129

Figure 5.1 Syntenic regions..145

Figure 5.2 Bias of mt-N genes in human, platypus and chicken....................147

Figure 6.1 Stability of DUF1220 domain counts..162

Figure 6.2 Distribution of likelihoods..163

Figure 6.3 Distribution of likelihoods..163

Figure 6.4 Plot of scaled-branch lengths from the variable rates model 164

Figure 6.5 Posterior means of DUF1220...169

Figure 6.6 Square root and *log* transformed DUF1220 associations.............171

Table of Tables

Table 2.1 Comparison of gene expression in spleen, heart, liver and gonad .. 58

Table 2.2 Comparison between the *log2* expression for male/female autosomal and Z expression. 58

Table 2.3 Contingency tables for all four tissues (-0.5-0.5) 69

Table 2.4 Contingency tables for all four tissues (-0.6-0.6) 69

Table 2.5 Contingency tables for all four tissues (strict ohnologs) 70

Table 2.6 Contingency tables for all four tissues (paralog comparison) 70

Table 2.7 Comparison of the degree of dosage compensation 74

Table 2.8 Significantly overrepresented GO terms 76

Table 3.1 Trinity assembly statistics for all six bird species. 89

Table 3.2 Assembly statistics 91

Table 3.3 Retained lineage-specific paralog pairs after each filtering step.... 102

Table 3.4 Pairs of Ensembl paralogs in family-structure filtered HOGs. 106

Table 4.1 Mean bias and 95% CI of mito-nuclear genes (1) 127

Table 4.2 Mean bias and 95% CI of mito-nuclear genes (2) 128

Table 4.3 Mean bias and 95% CI of mito-nuclear genes (3) 129

Table 5.1 Movement of mt-N genes on and off the X between 146

Table 5.2 Number of mt-N, total number of genes and the bias in distribution of mt-N genes on the human X and X-syntenic regions 148

Table 6.1 DUF1220 count data across 12 primates 161

Table 6.2 Phenotypic data using in the phylogenetic analyses 166

Table 6.3 MCMCglmm results of multivariate models 168

Table 6.4 MCMCglmm results using total genomic DUF1220 counts 170

Chapter 1

Introduction

Gene dosage

The genomic architecture of many different species has been revealed in increasing detail, largely due to the availability of modern sequencing technology. Genomes are not static, they undergo a wide range of large- and small-scale structural changes, including Whole Genome Duplications (WGD), chromosomal duplications (aneuploidies), Single Gene Duplications (SGD) and other Copy Number Variations (CNV). Even though genomic elements have been known to duplicate since the early days of genetics (Bridges 1936), the idea that these duplications are a major source of evolutionary novelty was first popularized by Susumu Ohno (Ohno 1970). Ohno proposed that beside the known mutational processes, the duplication of genes or whole genomes is a major source of variation during the evolution of organisms. In Ohno's view, evolutionary adaptations are often the result of using the 'raw material' generated by duplications. Thus, evolution tinkers often with duplicated genetic information rather than new genes (Jacob 1977).

A consequence of all genomic duplication or loss events is variation in gene dose, which describes the number of copies of a gene present in the genome. These gene dose changes can have a range of effects on a phenotype, caused by changing the activity level or gene expression of the duplicated genes. In this thesis, I investigate the consequences and evolutionary responses of gene dosage changes on gene expression.

Genome evolution

Duplications of genes or chromosomes have been known for some time; for example, the duplication of the *Drosophila* bar gene through unequal crossing over (Bridges 1936; Sturtevant 1925) or the chromosomal duplications in Jimson weed (*Datura Stramonium*) (Blakeslee, et al. 1920). The impact of such duplications on the phenotype and new functions was discussed by the founders of modern genetics (Haldane 1933; Muller 1935). By the 1960s the significance of duplications for the evolution of new functions was widely recognised (Nei 1969; Ohno, et al. 1968). Ohno proposed that mutations in functional genes are unlikely to be fixed in a population if they impair the current function (Ohno 1970). These selective constraints make it unlikely for a gene to evolve a new function, which often requires multiple mutations (Ohno 1970). As a possible solution, Ohno proposed that duplication events could resolve these constraints by creating new ‘raw material’. The new copies would create redundancy and selection would be free to shift one of the copies towards a new or specialised function. In the following introduction, I will discuss the different types of duplication events, some of the main theoretical models for the evolution after gene duplication and how selection can affect the distribution of genes in the genome.

Whole genome duplications

Whole Genome Duplication (WGD) events initially double the number of chromosomes in the genome. Ancient polyploidization events are those which occurred several million years ago and species that descended from these lineages are also known as paleopolyploid. The frequency of these events varies among lineages and many species have lost large amounts of the duplicated genetic material, making the detection of WGDs increasingly difficult over time. Recent advances in genome sequencing have ameliorated many of the original issues and facilitated the detection of WGD events.

In flowering plants, the estimated percentage of paleopolyploid species ranges from 30% to 70% (Soltis, et al. 2015). For example, the lineages leading to the wild ancestors of some important crop species, such as cereals, underwent polyploidization events (Paterson, et al. 2004). WGDs were also detected in well studied ‘model’ organisms, including *Zea mays* ca. 11 million years ago (mya) (Gaut, et al. 2000) and *Arabidopsis thaliana* ca. 38mya (Ermolaeva, et al. 2003; The Arabidopsis Genome Initiative 2000). In fungi, the availability of more than 40 different fully sequenced yeast genomes has made it possible to study the evolutionary dynamics of Single Gene Duplications (SGD) and WGDs on a large scale (Dujon 2010). Wolfe and Shields (1997) provided the first molecular evidence for an ancient WGD in baker’s yeast (*Saccharomyces cerevisiae*), which was subsequently confirmed by Kellis, et al. (2004). The polyploidization event took place ~150mya (Sugino and Innan 2005) and is possibly the result of an ancient interspecies hybridization event (Marcet-Houben and Gabaldón 2015). However, evidence for WGD events in other fungi outside the yeast lineage is still limited (Albertin and Marullo 2012).

Ohno hypothesised that the ancestor of all vertebrates underwent at least one, but more likely two rounds of WGD (The 2R hypothesis) (Ohno 1970). The 2R hypothesis was based on genome size and karyotype analyses and remained contentious for many years (Makałowski 2001). Initial analyses based on gene families, such as the *hox gene* cluster, seemed to confirm a pattern of four clusters of gene duplicates, which is consistent with the 2R hypothesis (Larhammar, et al. 2002; Lemons and McGinnis 2006). However, the sequencing of the human genome (Lander, et al. 2001; Venter, et al. 2001) revealed a lower number of protein coding genes than expected, and the tree topologies of human gene families with four members did not always show patterns that were consistent with the 2R hypothesis (Friedman and Hughes 2001). Testing the 2R hypothesis was hampered further by the ancient age and short succession of the WGD events around 450-500mya (Blomme, et al. 2006; Dehal and Boore 2005).

The analysis of paralogons, stretches of the genome containing gene duplicates in preserved order (Popovici, et al. 2001), suggested that a high number of gene duplications occurred in early chordates (McLysaght, et al. 2002), supporting at least one round of WGD. A similar analysis using paralogons and gene trees conducted by Dehal and Boore (2005) provided strong support for the 2R hypothesis, and by now the hypothesis has been widely accepted. In honour of Ohno's pioneering work genes duplicated in whole genome duplications have been termed 'ohnologs' or 'ohnologous' (Wolfe 2000; Wolfe 2001). In addition to the two rounds of WGD in the ancestor of modern vertebrates, several other WGD events occurred during the evolution of animals, such as in the ancestor of teleost fish ca. 350mya (Aparicio, et al. 2002; Jaillon, et al. 2004; Sato and Nishida 2010), or the clawed frog lineage (Uno, et al. 2013).

Evolution after WGD events

The retention of duplicated genetic material following polyploidization events is rare compared to SGDs, possibly because the initial duplication event is often lethal (Van de Peer, et al. 2009). However, in those lineages with one or more WGDs, duplicates can play an important role in adaptation to new environments, as the additional gene duplicates are under less constraint and are free to acquire new functions. In combination with higher functional redundancy and increased mutational robustness, the risk of extinction could be reduced (Crow and Wagner 2006). This view is supported by correlations between ancient WGD events and species radiations, for example in teleost fish (Blomme, et al. 2006; Sato and Nishida 2010) and flowering plants (Soltis, et al. 2009; Soltis, et al. 2015). WGDs could also facilitate speciation if the different gene pairs are lost in separate populations, potentially leading to reproductive isolation (Scannell, et al. 2006). However, clade age rather than WGDs may be more important for explaining species divergence (McPeek and Brown 2007). For example, invertebrates have diverged at similar rates compared to vertebrates, without any observed WGD events (McPeek and Brown 2007; Van de Peer, et al. 2009). This indicates that WGDs may facilitate divergence and increased speciation rates, but that they are not a necessary prerequisite.

In many instances, WGD events are followed by rapid and large-scale loss of many gene duplicates, as observed in teleost fish (Brunet, et al. 2006). In some cases, however, the loss occurs in a more gradual fashion over a longer time scale. For example, in *Paramecium* the stoichiometric constraints between gene duplicates may have led to the retention of many gene duplicates over a longer period of time (Aury, et al. 2006). The retention or loss of ohnolog pairs is non-random and depends on a number of different factors. Many models for the retention of single gene duplicates also apply to WGD events, and the neofunctionalisation or subfunctionalisation of recent duplicates can lead to the retention of those gene pairs. In contrast to small-scale duplications, WGD events only increase the absolute gene dosage: all genes are duplicated at the same time. As a result, in contrast to SGDs, the relative gene dosage does not change. For example, if two different genes (A and B) produce the two parts of a protein complex, a WGD event would double the gene dose of both genes (2A and 2B), but the stoichiometric balance would remain the same. While there is an absolute change, there is no change in relative gene dosage (2A divided by 2B = 1A divided by 1B) (Birchler, et al. 2005). This fact has important implication for the retention of dosage-sensitive genes. Genes are often part of tightly regulated networks and changes in gene dosage may disturb these networks (Birchler, et al. 2001; Veitia 2002).

Chromosomal aneuploidies, such as Turner syndrome (a missing X chromosome in females) (Turner 1938) or Down's syndrome (trisomy 21) (Lejeune, et al. 1959; Lejeune, et al. 1957), are well-described and have a significant phenotypic impact. The deletion of smaller numbers of genes can also lead to detrimental effect through negative gene dosage effects. For example, the deletion of 1.6mb on human chromosome 7 causes Williams-Beuren syndrome (Francke 1999), other Copy Number Variations (CNVs) have been associated with diseases, such as cancer (Shlien and Malkin 2009) and autism (Davis, et al. 2014). CNVs, include both deletions and duplications of smaller regions of the genome and are common in human populations (Feuk, et al. 2006; Redon, et al. 2006). A recent meta-study produced a detailed CNV map of the human genome and found that ~4.8-9.5% show variability in copy number (Zarrei, et al. 2015). However, the proportion of CNVs varies between chromosomes and depends on the genomic region. For example, regions close to the centromere or the telomere contain increased number of CNVs

(Nguyen, et al. 2006; Redon, et al. 2006; Zarrei, et al. 2015). Additionally, regions in vertebrate genomes containing ohnologs are underrepresented in CNVs, possibly as a result of negative dose effects (Makino, et al. 2013).

Not all genes, however, are haploinsufficient or show dosage effects. Zarrei, et al. (2015) found that around 100 genes can be deleted from the human genome without any apparent phenotypic effects. In yeast (*Saccharomyces cerevisiae*), a genome-wide scan revealed that only 3% of all genes are haploinsufficient (Deutschbauer, et al. 2005), many of which encode subunits of large protein complexes (Papp, et al. 2003). The Gene Dosage Balance Hypothesis (GBDH) predicts that dosage effects are a consequence of genes not acting in isolation (Papp, et al. 2003), instead requiring the co-expression of other genes in the genome to function correctly (Birchler, et al. 2005; Birchler and Veitia 2012; Veitia 2002). An increase in gene dose for some parts of these networks through single gene duplications may impair the functionality of other genes with detrimental fitness effects (Papp, et al. 2003). In a WGD, the relative dosage of all genes stays the same, which would shield dosage-sensitive genes against dosage imbalance. A subsequent loss would also be selected against because it would also result in an imbalance. Therefore, the GBDH predicts that dosage-sensitive genes should be overrepresented among genes retained in WGD events.

The GBDH hypothesis has been investigated in a range of organisms (Blanc and Wolfe 2004; Blomme, et al. 2006; Makino and McLysaght 2010; Papp, et al. 2003; Seoighe and Gehring 2004) and is well established (see Birchler and Veitia (2012) for an extensive review). If dosage-sensitive genes are more likely to be retained after a WGD event due to dosage constraints, these genes (ohnologs) can be used as a proxy for dosage-sensitivity. In line with these predictions, CNVs of ohnologs often have detrimental fitness effects (McLysaght, et al. 2014), and the phenotypic effects of Down’s syndrome could be due to the increased gene dosage of ohnologs on chromosome 21 (Makino and McLysaght 2010).

Detection of WGD events

The detection of WGD events has changed rapidly over the last years. Early work on WGDs was based on cytological evidence and genome size comparisons (Ohno 1970), or the analysis of specific gene family duplication structure, such as in the *hox* gene family (Larhammar, et al. 2002; Lemons and McGinnis 2006). In the absence of draft sequences of full genomes, this work remained limited, as only smaller subsets of genes could be investigated. The availability of modern sequencing technologies has helped tremendously to alleviate this issue and made it possible to investigate WGDs using whole genome comparisons.

The detection of WGDs is often a two-step procedure: In the first step, the sequence similarity between all genes within a genome is calculated using tools, such as Basic Local Alignment Search Tool (BLAST) (Altschul, et al. 1990), which performs a local alignment to calculate sequence similarity scores. These sequence similarity scores are then used to find homologous sequences that originated in a duplication event within that species (paralogs). Paralogs are subsequently clustered using methods such as single-linkage clustering (Dehal and Boore 2005). At this stage, however, it is not clear if the groups of paralogs originated in multiple single gene duplications or in a WGD. In order to differentiate between these alternatives, positional genomic information is used that can reveal patterns of conserved gene order between paralogs (syntenic regions). Conservation of gene order is indicative of large scale duplications. The detection of syntenic regions has been successfully applied to detect ohnologs in species such as yeast (Kellis, et al. 2004) and vertebrates (Dehal and Boore 2005; Makino and McLysaght 2010; Singh, et al. 2015).

For recent WGDs, strict positional information is often enough to recover polyploidization events, but WGDs are often followed by large-scale genomic rearrangements and massive gene loss. This complicates the detection of syntenic regions, especially for ancient WGDs. The two rounds of WGD in the common ancestor of all vertebrates occurred ~450-500mya, which makes the application of micro-synteny and strictly conserved gene order more challenging (Dehal and Boore 2005). An alternative approach is to use a wide sliding window that can span several hundred genes, which enables the detection of larger paralogous chromosomal regions (paralogons) (McLysaght, et al. 2002; Popovici, et al. 2001). Dehal and Boore (2005) used a relaxed synteny approach, which classified ca. 25% of the human genome as paralogons. This method depends on many *a priori* defined parameters, such as the window size, and it is unclear how to set these parameters without the availability of statistical measures to assess their impact on the discovery of paralogons. Consequently, different sets of genes were classified as ohnologs by different approaches, with little overlap between sets (Singh, et al. 2015).

The detection of WGDs often depends on the availability of at least one sequenced outgroup species that did not undergo a WGD. The outgroup species can be used as an anchor because for each single ortholog in the outgroup, at least two gene copies exist initially after a WGD in the species being investigated. For example, the discovery of the WGD event in *Saccharomyces cerevisiae* was facilitated by comparing the entire genome to the related yeast species *Kluyveromyces waltii* that diverged before the WGD event (Kellis, et al. 2004). Kellis, et al. (2004) found that for almost all genomic regions in *K. waltii* two copies were present in *S. cerevisiae*, providing strong support for an ancient genome duplication. However, the selection of the outgroup species may also introduce a bias to the analysis. If only one outgroup species is used, the annotation quality and genome build can influence the detection of ohnologs (Makino and McLysaght 2010; Singh, et al. 2015). This problem is often mitigated by using multiple outgroups and restricting the analysis to those genes that have support from more than one outgroup. The recent establishment of OhnologsDB (http://ohnologs.curie.fr/) has made sets of ohnologs available for many vertebrate taxa, with the additional benefit of a statistical measure indicating the confidence in the detection of the groups of ohnologs (Singh, et al. 2015). Singh, et al. (2015) used a synteny based approach that combines the comparison of multiple vertebrate species against a range of outgroups with self-comparisons of genomes.

Gene dosage and sex chromosome evolution

Sex chromosomes have evolved independently numerous times in different plant and animal lineages (Charlesworth 1991; Charlesworth 2013; Ellegren 2011). These cases of convergent evolution have led to a large diversity of sex chromosome systems (Bachtrog, et al. 2014; Beukeboom and Perrin 2014), and a large body of theory describes their evolution (Bull 1983; Charlesworth 1996; Charlesworth 1991; Charlesworth 2013; Charlesworth, et al. 2005; Rice 1984). Three major sex chromosome systems are known today, the male heterogametic XY system, the female heterogametic (ZW) system and the UV system in organisms where sex is expressed only in the haploid phase. In the context of genomic evolution, heteromorphic sex chromosomes provide a unique opportunity to understand how changes in gene dose for large numbers of genes impact the phenotype. The degeneration of the Y or W chromosome reduces gene dose in the heterogametic sex, resulting in unequal gene dose between the sexes. Viewed in this light, heteromorphic sex chromosomes are similar to aneuploidies, such as trisomy 21 in humans (Lejeune, et al. 1959; Lejeune, et al. 1957), which often have detrimental fitness effects.

The evolution of sex chromosomes is thought to occur in a multistep process, which ultimately results in the diverse range of sex chromosome systems we observe among extant taxa (Charlesworth, et al. 2005). One year after Sturtevant (1913) produced the first genetic map for *Drosophila*, Bridges (1914) showed that sex-linked genes are located on the *Drosophila* X chromosome, that the Y chromosome appears to be depleted of genes, and that there is no crossing over between the Y and X. This, in turn, inspired Muller (1914) to speculate that sex chromosomes could have evolved from regular autosomes and that the degradation of the Y chromosome is a gradual process. Muller (1914) reasoned that this process would be caused by the accumulation of recessive mutations, and the loss of genes will contribute to the Y chromosome degradation over time.

If dosage-sensitive genes are located on the sex chromosomes, the loss of one copy would result in negative fitness effects in the heterogametic sex. Ohno described this effect as the ‘peril of hemizygosity’ and proposed that a dosage compensation mechanism would be selected for to balance out these differences (Ohno 1967). This hypothetical mechanism of dosage compensation would restore the gene expression levels back to the ancestral level to balance the expression between the autosomes and the sex chromosomes, and between males and females. Balancing gene expression levels between males and females and between the autosomes and the sex chromosomes can be accomplished in several ways, and different mechanisms have been identified. Ohno’s discovery that the mammalian Barr body is an inactivated X chromosome (Ohno, et al. 1959) led him to formulate his classic two-step model of dosage compensation. First, the degradation of the Y or W chromosome leads to reduced gene dose and thus reduced gene expression in the heterogametic sex, which results in selection for hyperexpression of the remaining X or Z chromosome. This, in turn, causes overexpression in the homogametic sex, as gene expression is correlated between the sexes (Dean, et al. 2015). The inactivation of the X chromosome would counteract the overexpression in females and would balance out the gene dosage.

In *Drosophila melanogaster*, dosage compensation in males results from the hyperexpression of the X chromosome (Baker and Belote 1983; Conrad and Akhtar 2012). Unlike in mammals, hyperexpression in males is linked to the *male-specific lethal 2* (*MSL2*) gene and does not affect females, making X inactivation in females unnecessary (Conrad and Akhtar 2012). A dosage compensation mechanism analogous to the mammalian two-step process was found in *Caenorhabditis elegans*, where the hyperexpression of the single X chromosome in males is counteracted in hermaphrodites by the hypoexpression of both X chromosomes (Meyer 2010). All the described mechanisms have a key feature in common; the compensation mechanism affects large parts of the sex chromosomes and are therefore also referred to as ‘global or ‘complete’ dosage compensation mechanisms. The discovery of complete dosage compensation mechanisms in therian mammals, *Drosophila* and *C. elegans* seemingly confirmed Ohno's predictions and established the paradigm that the evolution of heteromorphic sex chromosomes is always associated with the evolution of a complete dosage compensation mechanisms (Mank 2013).

In 2007, two independent publications challenged this consensus by reporting the absence of a complete dosage compensation mechanism in *Gallus gallus* (Ellegren, et al. 2007; Itoh, et al. 2007). All birds possess a female heterogametic ZW system, with two ZZ chromosomes in males and a ZW configuration in females. Using microarray data, Ellegren, et al. (2007) and Itoh, et al. (2007) showed that the expression of the single Z chromosome in females is significantly lower than the autosomes, which is inconsistent with a complete dosage compensation mechanism. Some genes, however, still showed similar expression between males and females, suggesting a gene-by-gene regulatory mechanism. It is now clear that incomplete dosage compensation is present in other bird species (Naurin, et al. 2011; Uebbing, et al. 2013; Wolf and Bryk 2011), snakes (Vicoso, et al. 2013), fish (Chen, et al. 2014; Leder, et al. 2010) and some Dipterans (Vicoso and Bachtrog 2015).

Dosage compensation mechanisms are only necessary if dosage-sensitive genes are present on the sex chromosomes. For many genes on the sex chromosomes, halving the gene dose does not actually produce differences in gene expression (Malone, et al. 2012). Consequently, it may be enough to balance the expression of those genes that show negative dose effects, which would result in gene-by-gene dosage compensation (Mank and Ellegren 2008). The increasing availability of good-quality avian genomes in general, and especially the chicken, a model system for species with incomplete dosage compensation, offer an excellent opportunity to investigate if dosage-sensitive genes are dosage-compensated on a gene-by-gene basis.

Paralog divergence

Much theoretical and empirical work has been done to describe how single gene duplications impact the evolution of organisms. Gene duplications can arise through a range of different mechanisms. Large-scale events, such as WGD events (see above) and chromosomal duplications play an important role, but other mechanisms also contribute to the duplication on a smaller scale. Unequal cross over, where homologous chromosomes exchange unequal parts of DNA during recombination, result in tandem duplications of genes on one chromosome and deletion on the other (Hurles 2004). Tandem duplications generated by unequal crossover are thought to be one of the primary mechanisms creating single gene duplications and, for example, account for around 80% of all lineage-specific gene duplicates in *Drosophila melanogaster* or *Drosophila yakuba* (Zhou, et al. 2008). However, other mechanisms, such as retrotransposition may contribute a significant fraction of gene duplicates. During retrotransposition, processed and spliced mRNA is transcribed back into cDNA and re-introduced into the genomic sequence. These retrotransposed genes lack the intron-exon structure of other genes and are usually pseudogenes as they lack the proper regulatory sites for correct expression (Kaessmann, et al. 2009). Nevertheless, a minority of retrotransposed genes can become active and provide novel, functional duplicates (Vinckenbosch, et al. 2006).

Several models have been proposed which seek to describe the evolutionary dynamics after gene duplication (reviewed in Innan and Kondrashov 2010; Taylor and Raes 2004; Zhang 2003). Broadly, these models can be divided into two categories: those that assume a phase of neutral evolution after duplication, and those that state that duplication may have an immediate selective advantage (Innan and Kondrashov 2010). In all cases, a duplication event generates two copies of the same gene (Ohno 1970; Zhang 2003). Regardless of the underlying duplication mechanism, gene duplicates are termed *paralogs*. It is also assumed that gene duplication leads to the creation of two functionally indistinguishable copies.

In the first category of models, the generation of two copies in the genome is assumed to be initially neutral (Ohno 1970), and the fixation probability of both copies in a diploid population is 1/2N (N= population size) (Kimura 1984), which is relatively small for species with a large population size. Under neutrality, mutations can accumulate in one copy while the other copy maintains the ancestral function. However, mutations are much more likely to be deleterious than beneficial, as functionally advantageous changes often require multiple mutations (Nei 1969). One copy may therefore accumulate deleterious mutations faster than selection can act to produce a new function. This can then lead to pseudogenisation through frameshift mutations or the introduction of a premature stop-codon. The non-functional copy may subsequently be lost through random genetic drift (Kimura and King 1979), the predominant fate of duplicated genes (Lynch and Conery 2000; Zhang 2003).

The first adaptive alternative to pseudogenisation and subsequent loss is known as the neofunctionalisation model (Force, et al. 1999; Ohno 1970). After duplication, functional divergence between paralogs can occur when one copy acquires mutations that shift the function of a gene in a new direction. Ohno realised that this scenario is less likely than the accumulation of detrimental mutations; however, in those cases where beneficial mutations accumulate, selection can act to fix this variant in the population. This functional divergence of paralogs is dependent on the maintenance of one copy that is still able to carry out the original function. Many examples of neofunctionalisation have been described; for example, Assis and Bachtrog (2013) found that the retention of nearly two-thirds of young duplicates in *Drosophila* can be explained by the neofunctionalisation model. In a number of different yeast species, ohnologs (paralogs generated via a WGD event) also showed evidence of neofunctionalisation after a duplication event (Byrne and Wolfe 2007). In vertebrates, examples include the neofunctionalisation of retinoic acid receptors (Escriva, et al. 2006), or the independent duplication of a pancreatic ribonuclease gene in Asian and African leaf monkeys (Zhang 2006; Zhang, et al. 2002).

An alternative fate for paralogs is described by the subfunctionalisation or duplication-degeneration-complementation model (DDC) (Force, et al. 1999; Ohno 1970; Stoltzfus 1999). The DDC model is based on the realisation that a loss of function mutation is unlikely to eliminate all aspects of gene function at once, as genes can fulfil several different subfunctions by being actively expressed in a variety of tissues and at different points in time (Lynch and Force 2000). This model is also consistent with the observation that many genes contain multiple protein domains, with different functions that can evolve as modular units (Bornberg-Bauer and Albà 2013). The DDC model proposes that under neutral expectations paralogs can accumulate complementary degenerative mutations that partition the functionality between the copies (Force, et al. 1999; Stoltzfus 1999). The subfunctionalisation of the paralogs means both copies evolve under selection to maintain the ancestral gene’s function (Lynch and Force 2000). The DDC model does not conflict with the neofunctionalisation and nonfunctionalisation model, but it does challenge the notion that gene duplication results either in loss or the evolution of a new function.

A third outcome for paralogs is described by a model known as Escape from Adaptive Conflict (EAC) (Des Marais and Rausher 2008; Hughes 1994; Piatigorsky 1991). The EAC is similar to the DDC model as it also describes the subfunctionalisation of the duplicates, but it assumes adaptive evolution follows a duplication event. A gene with one major function may have a second minor function. Prior to the duplication event, the secondary function may evolve under pleiotropic constraint, as adaptive mutations for the secondary function would have detrimental effects on the main function (Hughes 1994; Piatigorsky 1991). A duplication event removes this constraint, providing the opportunity for selection to favour the accumulation of adaptive mutations that improve the alternative function in one copy. However, the EAC model is difficult to distinguish from both the neofunctionalisation or the subfunctionalisation model, even though many cases that were described as neofunctionalisation may better fit the EAC model (Des Marais and Rausher 2008). Des Marais and Rausher (2008) proposed that one way of distinguishing the EAC model from the neofunctionalisation model is to search for adaptive mutations in both copies. Under the neofunctionalisation model adaptive mutations are only predicted to accumulate in one copy.

The second class of gene duplication models acknowledges the fact that a duplication could be immediately beneficial. The duplication of a single gene increases the gene dosage, and subsequently could increase the gene expression level and therefore gene product. In contrast to all previous models, which assume a period of neutral divergence after duplication, gene dosage benefits offer an explanation for how duplication events can be immediately beneficial (Kondrashov, et al. 2002). The duplicate could then be fixed by natural selection. For example, an increased copy number of the gene *pfmdr1* in *Plasmodium falciparum* increases the resistance of *P. falciparum* against mefloquine, a major antimalarial drug (Price, et al. 2004). Increased gene copy number (also called 'gene amplification') is common in bacteria, where it is considered a major avenue for adaptive evolution (Andersson and Hughes 2009). In human populations, beneficial gene dose changes are known for genes encoding immunity related proteins, such as the duplication of the *CCL3L1*, where an increase in copy number is related to a lower susceptibility for HIV/AIDS (Gonzalez, et al. 2005). Another example is the increase in copy number of salivary amylase genes in humans, which correlates with the ability to digest starch rich food (Perry, et al. 2007). This is consistent with a dietary shift during hominid evolution from low-starch to high-starch diets (Perry, et al. 2007).

Protein domain evolution

In addition to affecting whole genomes, chromosomes or single genes, duplication and loss events can also affect segments of genes. For coding genes, the translated protein often consists of one or more independent folding units or domains (Buljan and Bateman 2009; Coulson and Moult 2002; Rossmann, et al. 1974). These structural domains enable the protein to carry out its function for example by providing a binding interface to DNA or RNA. Most genes contain more than one domain, which results in a specific protein domain arrangement (Björklund, et al. 2005). The number of known protein domains is stabilising with ca. 15,000 listed in the PFAM database (Finn, et al. 2014). In contrast, the number of different domain arrangements continues to increase rapidly, with more than 75,000 having been identified (Levitt 2009). Domain arrangements evolve by the duplication, loss or rearrangement of domains, which results in a form of modular protein evolution (Björklund, et al. 2005; Bornberg-Bauer, et al. 2005; Buljan and Bateman 2009; Moore, et al. 2008). By re-using already existing domains in a different context, proteins can acquire new functions and potentially facilitate or contribute to the evolution of organismal complexity (Vogel and Chothia 2006).

Many proteins contain repeats of the same domain (Björklund, et al. 2006) and these repeats evolve rapidly through internal tandem duplications of single domains or groups of domains (Björklund, et al. 2006; Björklund, et al. 2010). Gain of domains through duplication can be interpreted as an increase in domain dose in an analogous way described above for gene dose. Domain repeats are also overrepresented among genes involved in segmental duplications (Björklund, et al. 2010), and variation in domain repeat number can be observed at a population level (Bornberg-Bauer and Albà 2013). Domain repeats have been associated with adaptive processes in *Saccharomyces cerevisiae*, where intragenic repeats increase the diversity of surface antigens (Verstrepen, et al. 2005), demonstrating their potential to contribute to organismal fitness.

The expansion of some protein domain repeats, such as the zinc-finger transcription factors (Emerson and Thomas 2009) or the DUF1220 family (Popesco, et al. 2006), provide examples of lineage-specific amplification, which could have played a crucial role during human evolution. Investigating lineage-specific amplification of protein domain repeats remains challenging because they depend on robust methods to assess domain counts in a phylogenetic framework. BLAST (Altschul, et al. 1990) based methods can introduce a phylogenetic bias when query sequences are used from a single species. Instead, these methods can be replaced by more appropriate tools, such as Hidden Markov Models (HMMs). HMMs are able to identify protein domains by matching a sequence to a probabilistic model of the protein domain sequence structure. HMMs are readily available in the PFAM database (Finn, et al. 2014) and the availability of the HMMER3 toolkit allow HMM based protein domain searches at speeds comparable to BLAST (Eddy 2011; Wheeler and Eddy 2013). Additionally, robust protein domain annotation needs to be combined with comparative phenotypic data, such as brain volume data across primates, to understand how an increase in protein domain dose affects the phenotype.

Gene movement and rearrangements

A consequence of duplication and loss events is that the genomic landscape changes over time. The location and distribution of genes are especially important when genes interact with each other to produce complex phenotypes. In these cases, selection may favour clusters of tightly linked genes that are more likely to be inherited together. For example, this is the case for sex chromosomes (Charlesworth 1991) or supergenes (Schwander, et al. 2014), which lack recombination. In contrast, other regions of the genome show elevated recombination rates. This can result in the relatively frequent generation of duplications through unequal crossing over (Hurles 2004) and non-allelic homologous recombination (Redon, et al. 2006). Genes located in regions with elevated recombination rate show frequent copy number variations, as is the case for the DUF1220 protein domains (Keeney, et al. 2014).

Single gene duplication events, followed by loss of one copy, can lead to gene movement from one region in the genome to another. In some cases, selection may favour the new location over the old one, which can result in the underrepresentation of classes of genes in specific regions of the genome. For example, in *Drosophila* a large number of genes with male-related function have moved off the X chromosome (Meisel, et al. 2009; Vibranovski, et al. 2009), potentially because selection favoured a location where these genes would escape X inactivation during spermatogenesis (Betrán, et al. 2002), or as a potential escape from sexual antagonism (Rice 1984).

The genomic location of genes is also important for the correct function of the mitochondria. Eukaryotic cells contain both the nuclear and the mitochondrial genome, and many mitochondrial genes moved to the nuclear genome (Gillham 1994). Consequently, many gene products necessary for mitochondrial function are encoded in the nuclear genome (called mito-nuclear genes or mt-N) and subsequently exported to the mitochondria. Mito-nuclear genes are highly conserved (Gillham 1994), and compatibility between the nucleus and the mitochondrion is crucial to the fitness of eukaryotic organisms (Meiklejohn, et al. 2013). In contrast to the autosomes, the mitochondrial genome is transmitted almost exclusively through the matriline, and genes located on the mitochondrial genome are selected for female fitness, as mitochondria in males will not be passed on to the next generation (Unckless and Herren 2009; Zhang, et al. 2012). This can result in the accumulation of mutations that are detrimental to males but which have no negative fitness effects on females. As a consequence, some male-specific genetic disorders are caused by genes located on the mitochondrial genome (Taylor and Turnbull 2005), a pattern sometimes referred to as the ‘mother’s curse’ (e.g. Gemmell, et al. 2004).

Summary of aims

In this thesis, I report the results of a series of studies that investigate how changes in the genomic architecture influence the evolution of organisms. Specifically, I aim to understand:

- How does the evolution of heteromorphic sex chromosomes affect the gene expression patterns of dosage-sensitive genes located on the sex chromosomes?

- Can RNA-Seq based *de novo* assemblies be used to detect lineage-specific paralogs, and how does the gene expression of those paralogs evolve?

- How does the matrilineal descent of mitochondria affect the distribution of genes that interact with the mitochondria?

- What are the phenotypic effects of a lineage-specific increase in protein domain dose?

Summary of thesis chapters

In **chapter 2**, I analyse RNA-Seq data from four different chicken (*Gallus gallus*) tissues in both males and females to analyse dosage compensation on the sex chromosomes. I use ohnologs from ancient WGD as proxies for dosage sensitivity and show that ohnologs located on the chicken Z chromosome are preferentially dosage compensated.

In **chapter 3**, I use *de novo* transcriptomes from RNA-Seq data from six different bird species to detect lineage-specific paralogs in order to test how gene expression diverges in recent gene duplicates.

In **chapter 4**, I investigate the genomic distribution of mito-nuclear genes in the genomes of a range of organisms with different sex chromosome systems. I use these data to understand whether the different inheritance patterns of sex chromosomes and the mitochondria result in sexual conflict over mito-nuclear genes and analyse whether this causes mito-nuclear genes to move off the sex chromosomes.

In **chapter 5**, I use synteny information to understand the evolution of the mammalian X chromosome from a pair of autosomes. I investigate whether the underrepresentation of mito-nuclear genes on the X chromosome is a function of X chromosome evolution, or if the paucity of genes was just a chance event.

In **chapter 6**, I investigate the copy number increase of DUF1220 protein domains in the primate phylogeny and their impact on brain evolution. I use custom-built Hidden Markov Models to detect and count DUF1220 protein domains and use this count data to link it with phenotypic measures of brain component size.

Glossary

Aneuploidy An abnormal chromosomal copy number, either through loss or gain, compared to the normal chromosome number of a species.

Dosage sensitivity Genes where a change in dose, either through duplication or loss, has negative fitness effects.

Gene dosage/genetic dose The copy number of a gene in the genome.

Haploinsufficiency A gene that is unable to function correctly when the dose is reduced e.g. in a diploid organism one copy of the gene is lost or inactive and the remaining copy cannot produce enough gene product.

Heteromorphic sex chromosomes A sex chromosome system in which one chromosome is visibly degenerated in comparison to the other e.g. the mammalian XY system.

Homology (gene) Genes are homologous if they share a common ancestor.

Inparalog Genes duplicated *after* a defined speciation event (Sonnhammer and Koonin 2002).

Mito-nuclear gene Nuclear genes whose gene products are exported to the mitochondria.

Ohnolog Genes that were duplicated in a whole genome duplication (Wolfe 2000; Wolfe 2001).

Ortholog Genes related to each other through a speciation event (Fitch 1970).

Outparalog Genes duplicated *before* a defined speciation event (Sonnhammer and Koonin 2002).

Paralog Genes related to each other through a duplication event (Fitch 1970).

Polyploidy Species with multiple sets of homologous chromosomes. Can be specified as diploid (2 homologous chromosomes, triploid (three homologous chromosomes) tetraploid, etc.

Abbreviations

ASE	Allele-Specific Expression
BLAST	Basic Local Alignment Tool
CI	Confidence Intervals
CNV	Copy Number Variation
CPM	Counts Per Million
DC	Dosage Compensation
DDC	Duplication-Degeneration-Complementation
EAC	Escape from Adaptive Conflict
GDBH	Gene Dosage Balance Hypothesis
GO term	Gene Ontology term
Mt-N	Mito-nuclear
(n)HMM	(nucleotide) Hidden Markov Model
rBBH	reciprocal Best BLAST Hit
RPKM	Reads Per Kilobase of transcript per Million mapped reads
SGD	Single Gene Duplication
SNP	Single Nucleotide Polymorphism
WGD	Whole Genome Duplication

Additional published work

In addition to the work presented in this thesis, I also contributed various analyses to a range of other published studies:

- Harrison PW, Wright AE, **Zimmer F**, Dean R, Montgomery SH, Pointer MA, Mank JE (2015) Sexual selection drives evolution and rapid turnover of male gene expression. *Proceedings of the National Academy of Sciences, USA* 112: 4393-4398 Copyright (2015) National Academy of Sciences.
  doi: 10.1073/pnas.1501339112

- Dean R, Harrison PW, Wright AE, **Zimmer F**, Mank JE (2015) Positive selection underlies Faster-Z evolution of gene expression in birds. *Molecular Biology and Evolution* 32: 2646-2656
  doi: 10.1093/molbev/msv138

- Wright AE, **Zimmer F**, Harrison PW, Mank JE (2015) Conservation of regional variation in sex-specific sex chromosome regulation. *Genetics* 201: 587-598
  doi: 10.1534/genetics.115.179234

- Wright AE, Harrison PW, **Zimmer F**, Montgomery SH, Pointer MA, Mank JE (2015) Variation in promiscuity and sexual selection drives avian rate of Faster-Z evolution. *Molecular Ecolog*y 24:1218-1235
  doi: 10.1111/mec.13113

- Wright AE, Dean R, **Zimmer F**, Mank JE (2016) How to make a sex chromosome. *Nature Communications* 7:12087
  doi: 10.1038/ncomms12087

- Schmitz JS, **Zimmer F**, Bornberg-Bauer E (2016) Mechanisms of transcription factor evolution in Metazoa. *Nucleic Acids Research* 13: 6287-6297
  doi: 10.1093/nar/gkw492

Reprints of these publications are included in the appendix.

Chapter 2

Compensation of dosage-sensitive genes on the chicken Z chromosome

The analyses presented in this chapter have been published in
Genome Biology and Evolution:

Zimmer F, Harrison PW, Dessimoz C, Mank JE (2016) Compensation of dosage-sensitive genes on the chicken Z chromosome. *Genome Biology and Evolution* 8 (4): 1233-1242, **doi**: 10.1093/gbe/evw075

Author contributions:

I designed the data analyses with Dr Peter Harrison and Professor Judith Mank and wrote the paper in collaboration with both of them.

Professor Dessimoz provided support in designing the ohnolog analyses.

Professor Judith Mank collected all samples and performed the RNA extractions.

Summary

In many diploid species, sex determination is linked to a pair of sex chromosomes that evolved from a pair of autosomes. In these organisms, the degeneration of the sex-limited Y or W chromosome causes a reduction in gene dose in the heterogametic sex for X- or Z-linked genes. Variations in gene dose are detrimental for large chromosomal regions when they span dosage-sensitive genes, and many organisms were thought to evolve complete mechanisms of dosage compensation to mitigate this. However, the recent realization that a wide variety of organisms lack complete mechanisms of sex chromosome dosage compensation has presented a perplexing question: How do organisms with incomplete dosage compensation avoid deleterious effects of gene dose differences between the sexes? Here, I use expression data from the chicken (*Gallus gallus*) to show that ohnologs, duplicated genes known to be dosage-sensitive, are preferentially dosage-compensated on the chicken Z chromosome. These results indicate that even in the absence of a complete and chromosome-wide dosage compensation mechanism, dosage-sensitive genes are effectively dosage-compensated on the Z chromosome.

Introduction

Heteromorphic sex chromosomes have evolved independently in many species (Bachtrog, et al. 2014; Beukeboom and Perrin 2014). In some cases, recombination has been suppressed along the majority of the length of the sex chromosomes, leading to a large-scale loss of active genes from the sex-limited Y and W chromosomes (Bachtrog, et al. 2011; Charlesworth, et al. 2005). This results in large differences in size, with one large, gene-rich chromosome (X or Z chromosome) and one smaller chromosome that lacks many genes (Y or W chromosome).

The decay of Y and W chromosome gene content leads to differences in gene dose between the sexes, where the heterogametic sex has one half of the dose of all genes lost from the sex-limited chromosome compared to the homogametic sex. For many loci, gene dose correlates with gene expression (Birchler, et al. 2005; Malone, et al. 2012; Pollack, et al. 2002; Torres, et al. 2007); therefore, the reduced gene dose on the X or Z chromosome should result in reduced gene expression in the heterogametic sex. When dosage-sensitive genes are affected, this could lead to a reduction in fitness in the heterogametic sex and result in selective pressures favouring the evolution of dosage compensation mechanisms (Charlesworth 1996, 1978, 1998; Ohno 1967). These mechanisms should equalize the expression between the sex chromosomes and the autosomes, thereby restoring them to the ancestral level before the evolution of sex chromosomes. Second, the mechanism should equalize the expression of individual dosage-sensitive genes between males and females.

Although it was once assumed that complete and global dosage compensation would always be associated with sex chromosome evolution (Ohno 1967), there is considerable variation in the mechanism and completeness of dosage compensation across species. For example, in *Drosophila melanogaster* (Conrad and Akhtar 2012) and *Caenorhabditis elegans* (Meyer 2010), dosage balance is achieved through regulatory mechanisms affecting the entire X chromosome (Straub and Becker 2007). In these cases, differences in gene dose of the sex chromosome are compensated for and expression is on average balanced between the sexes for the X chromosome. Additionally, the expression of the single X and the diploid autosomes in heterogametic males is balanced. However, it is now clear that complete mechanisms of dosage compensation are rare, and many organisms, including birds (Ellegren, et al. 2007; Itoh, et al. 2007; Naurin, et al. 2011; Uebbing, et al. 2013; Wolf and Bryk 2011; Wright, et al. 2015b), snakes (Vicoso, et al. 2013), many insects (Vicoso and Bachtrog 2015) and fish (Chen, et al. 2014; Leder, et al. 2010), have incomplete dosage compensation (reviewed in Mank 2013).

Incomplete dosage compensation was first documented in chicken (Ellegren, et al. 2007; Itoh, et al. 2007) and subsequently confirmed in several other avian species (Naurin, et al. 2011; Uebbing, et al. 2013; Wolf and Bryk 2011; Wright, et al. 2015b). In birds, which are a model for studies of incomplete dosage compensation, there is a significant reduction in average expression of the Z chromosomes in females, the heterogametic sex relative to the autosomes, as well as a reduction in the male Z chromosome average (Ellegren, et al. 2007; Itoh, et al. 2007; Uebbing, et al. 2015; Uebbing, et al. 2013; Wolf and Bryk 2011). The realisation that many organisms with heteromorphic sex chromosomes have not in fact evolved complete and global dosage compensation mechanisms is perplexing, as it is unclear how these organisms cope with negative dose effects. A reduction in gene dose often does not produce an observable difference in expression for many genes (Malone, et al. 2012), and it is unclear whether certain loci are actively dosage-compensated or simply lack dose effects.

One possible explanation proposed by Mank and Ellegren (2008) is that instead of requiring a global mechanism of dosage compensation, the regulation of gene dose might occur on a gene-by-gene basis. A more targeted, local mechanism of dosage compensation should primarily affect the expression of dosage-sensitive genes (Mank, et al. 2011). The role of dosage-sensitivity for the evolution of dosage compensation mechanisms has been discussed by a number of reviews (Ercan 2015; Mank 2013; Pessia, et al. 2013;

Veitia, et al. 2015) and has been investigated in a range of species. For example, in mammals X chromosomal expression is reduced compared to the autosomes in both males and females (Julien, et al. 2012; Xiong, et al. 2010), possibly as a consequence of X chromosome inactivation. However, dosage-sensitive genes, such as protein-complexes, show evidence of a higher degree of dosage-compensation (Lin, et al. 2012; Pessia, et al. 2012) compared to other gene categories. Recent studies in nematodes (Albritton, et al. 2014) and fish (White, et al. 2015) also showed similar patterns of compensated dosage-sensitive genes.

Dosage sensitivity can result from interactions with other genes or gene products (Veitia 2004), as in the case of transcription factors and large protein complexes (Papp, et al. 2003). Individual duplications of these dosage-sensitive genes are likely to be rare, as they disrupt the stoichiometric balance and may disturb gene networks (Birchler, et al. 2001; Birchler and Veitia 2012; Papp, et al. 2003). However, dosage-sensitive genes should be preferentially retained after Whole Genome Duplications (WGDs) (Birchler and Veitia 2012; Edger and Pires 2009; Papp, et al. 2003). In contrast, dosage-insensitive genes that do not exhibit neo- or sub-functionalisation are often lost after WGD (Dehal and Boore 2005). WGDs have occurred in a wide range of lineages (Cui, et al. 2006; Dehal and Boore 2005; Kellis, et al. 2004; Van de Peer, et al. 2009; Wolfe and Shields 1997), including two rounds of WGD events roughly 500 MYA ago (Dehal and Boore 2005), which gave rise to ca. 16%-34% of the chicken genome (Singh, et al. 2015).

Preferentially retained gene duplicates originating from WGDs, also known as ohnologs (Wolfe 2000; Wolfe 2001), are skewed towards gene families associated with dosage-sensitive functions such as signalling and development (Blomme, et al. 2006) and protein complexes (Makino, et al. 2009). The dosage sensitivity of ohnologs (Blomme, et al. 2006; Makino, et al. 2009) is well established and makes them particularly useful in assessing the effectiveness of incomplete dosage compensation. I therefore use ohnologs to investigate the effectiveness of compensation on the chicken Z chromosome and to understand the evolution of incomplete sex chromosome dosage compensation mechanisms in general.

Material and Methods

RNA-Seq analysis and gene expression estimates

Heart, liver and spleen samples from White Leghorn chicken (*Gallus gallus*) embryonic day 19 eggs incubated under standard conditions were collected. Embryos were sexed visually and based on expression of W-linked genes. For each tissue, four biological samples were collected for both males and females. One female liver sample was excluded from the analyses because it showed only spurious W expression and when investigating the Z:A ratio it was clearly masculinized. All samples were first stored in RNAlater (Qiagen) and then total RNA was extracted (Qiagen Animal Tissue RNA kit).

Library construction and Illumina sequencing was done at the Wellcome Trust Centre of Human Genetics (WTCHG), Oxford. Each sample was normalised to 2.5μg total RNA prior to a PolyA isolation using an NEB Magnetic mRNA Isolation Kit. PCR was carried out over 15 cycles using custom indexed primers (WTCHG). Libraries were quality controlled with picogreen and tapestation, and were subsequently normalised equimolarly into 12-plex pools for Illumina HiSeq sequencing. Heart, liver and spleen samples were sequenced using an Illumina HiSeq 2000 as paired-end 100-bp reads. 51-bp paired end reads of gonadal samples from the same development stage were obtained from (Moghadam, et al. 2012).

I trimmed each library using Trimmomatic v0.22 (Bolger, et al. 2014; Lohse, et al. 2012), removing leading and trailing bases with a Phred score < 4 and trimming using a sliding window approach when the average Phred score over four bases was less than 15. Reads were kept if they were at least 36 bases after trimming. Libraries were quality-inspected manually using FASTQC v0.10.1 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The trimmed libraries were aligned against the chicken reference genome Ensembl version 75 Galgal4 (Cunningham, et al. 2015) using TopHat v2.0.11 (Kim, et al.

2013) and bowtie2 v2.2.2 (Langmead and Salzberg 2012), allowing five mismatches to the reference genome with on average 17 million paired-end mappable reads per sample. Multi-mapping reads were removed and I then sorted and indexed the resulting alignment files for each library separately using Samtools v0.1.18/9 (Li, et al. 2009).

I extracted reads mapping to annotated genes using HTseq-Count v0.6.1p1 (Anders, et al. 2015) and normalised all tissues separately using the Trimmed Mean of M-values (TMM) method (Robinson and Oshlack 2010) available in edgeR v3.2.4 (Robinson, et al. 2010). I estimated differential expression between males and females in all tissues using edgeR‘s exactTest method and exported the log_2 fold change (log_2FC; female – male expression), average log_2 count per million (logCPM), FDR corrected *P*-values from the exactTest function and individual CPM (Counts Per Million) values for all samples and genes. Genes were only included when the average CPM was larger than 2 across all males and females, filtering out loci with low expression. When comparing groups of genes to each other, I normalised the CPM values by gene length, resulting in RPKM (Reads Per Kilobase of transcript per Million mapped reads) values. Only genes annotated to the autosomes and the Z chromosome were assessed. Individual genes were defined as dosage-compensated on the Z chromosome if the female:male log_2 fold change ranged from -0.5 to 0.5 (Wright, et al. 2015b). I defined genes as sex-biased if the edgeR exactTest was significant after FDR correction ($q < 0.05$) and the log_2 fold change was 1 > for female-biased genes or < -1 for male-biased genes.

Identification of ohnologs and other paralogs

I used the Ohnologs database (http://ohnologs.curie.fr/) (Singh, et al. 2015) to obtain ohnologs present in the chicken genome. I used the relaxed set of ohnologs as the primary dataset, in order to maximize the number of ohnologs. Additionally, I used the Ensembl REST API (accessed February 2015) (Yates, et al. 2015) to identify all paralogs in the chicken genome, which also includes those homologs originated in single-gene duplications.

Functional annotation of ohnologs

I used the G:profiler toolkit (Reimand, et al. 2011) to perform GO Term (Ashburner, et al. 2000) overrepresentation analyses. All ohnologs on the Z chromosome were provided as an input list and compared to the entire genomic background, using only genes with annotated GO terms in the comparison. Standard settings were used and GO Terms were only considered if they had a significant *P*-value after multiple testing correction via G:Profiler's G:SCS method (*P*-value < 0.05). Additionally, I used the CORUM database (Ruepp, et al. 2010), version from February 2012, to annotate protein complexes in the chicken genome. The CORUM database contains only mammalian data and I used the Ensembl REST API (Yates, et al. 2015) to detect the corresponding chicken homologs, where possible.

SNP calling and estimation of allele specific expression

In order to detect allele-specific expression (ASE) from RNA-Seq data, I modified a pipeline from Quinn et al. (Quinn, et al. 2014a; Quinn, et al. 2014b). As I was interested in detecting ASE on the Z chromosome, I only called Single Nucleotide Polymorphisms (SNPs) in the homogametic sex (males) for each tissue. SNPs were called using Samtools mpileup v0.1.18 (Li, et al. 2009) and VarScan2 v2.3.6 (Koboldt, et al. 2012). SNPs were called separately for each tissue using all four available male samples. I required minimum coverage of 2 and minimum Phred score of 20 (--min-avg-qual 20) to call a SNP and also required a minimum frequency of 0.9 to call a homozygote (--min-freq-for-hom 0.9). The resulting Variant Call Formatted (VCF) files were then filtered further to remove noise and increase SNP call confidence. In a first step, I filtered out SNPs using a combination of a fixed minimum threshold of 17 reads per site (the combination of major and minor allele) in all samples, as a power analysis indicates that a 17 read coverage for a SNP results in 73% power to detect allele specific-expression and also excluded all SNPs with more than two alleles. I also used a variable threshold that accounts for the likelihood of observing a second allele because of sequencing errors an error probability of 1 in 100 (Quinn, et al. 2014a) and a maximum coverage of 100,000. RNA-Seq data has an intrinsic bias for the estimation of ASE, because those reads that resemble the reference genome have a higher probability of aligning successfully. In order to remove this bias, I eliminated clusters of SNPs if there were more than 5 SNPs in a window of 100 base pairs (Stevenson, et al. 2013). I used BEDtools intersect v2.20.1 (Quinlan and Hall 2010) to filter out all SNPs that were not located in a known transcript.

If both chromosomes are active to the same degree, I expect that the probability of observing reads from one or the other chromosome is 0.5. I therefore used a two-tailed binomial test to show significant deviations from this expected distribution ($P < 0.05$). Binomial tests were corrected for multiple testing on the autosomes, because of the larger number of testable sites. In order to account for the fact that binomial tests will be significant even for very small deviations in the observed distribution when the sample size, in this case the alignment depth, is big enough, I also employed a minimum threshold of 70% reads stemming from one allele to call significant allele-specific expression. Additionally, I used a power analysis to ensure that the ability to detect ASE is sufficient. At a minimum coverage of 17 reads per site, the power for detecting ASE is greater than 73%, which suggests that I was able to detect patterns of ASE successfully in most cases. I only included genes in the analysis if at least one SNP showed consistent allele-specific expression across all samples.

All analyses and statistical comparisons were performed using Python, Matplotlib (Hunter 2007) and R (R Core Team 2015). Code and iPython notebooks (Pérez and Granger 2007) are available on GitHub at https://github.com/qfma/ohnolog-dc. All sequencing data used in the analyses is available in the NCBI Short Read Archive under accession number SRP065394.

Results

I generated RNA-Seq gene expression profiles from multiple male and female biological replicates for four different tissues (spleen, heart, liver and gonad) in chicken (*Gallus gallus*), recovering on average 17 million paired-end mappable reads per sample. I removed genes that were not expressed on average in all male and female above at least two Counts Per Million (CPM). The number of genes expressed on the autosomes and Z chromosome for each tissue are shown in **Table 2.1**.

Incomplete dosage compensation in females and reduced Z expression in males

Dosage compensation has been assessed in a variety of ways, often depending on the system being studied. I used two approaches to assess dosage compensation status. First, complete dosage compensation should equalize female Z-linked and autosomal expression. Second, dosage compensation can also act on a local gene-by-gene basis, balancing the individual gene expression in males and females, which may be the dominant mechanism for dosage-sensitive genes.

Consistent with previous studies showing the incomplete dosage compensation in chicken, I detected lower average expression of Z-linked genes in comparison to autosomal genes in all female tissues (spleen $P < 0.0001$, Z-score = 11.19; heart $P < 0.0001$, Z-score = 11.22; liver $P < 0.0001$, Z-score = 8.88; ovaries $P < 0.0001$, Z-score = 9.20; Wilcoxon Rank Sum Test; **Figure 2.1**, **Figure 2.2**, **Table 2.2**). It is also expected that the average expression of the Z chromosomes in males is similar to the autosomal average, as two Z chromosomes are present. In line with this prediction, the distribution of male expression is not significantly different to the autosomes in testes ($P = 0.79$, Z-score = 0.27, Wilcoxon Rank Sum Test). However, previous work has indicated that in some tissues, expression of the Z in males is also less than the autosomal average (Julien, et al. 2012), and I also recovered a significant reduction in average expression of Z-linked loci compared to average autosomal expression in all somatic tissues in males (spleen *P* < 0.0001, Z-score = 5.50; heart *P* < 0.0001, Z-score = 6.69; liver *P* < 0.0001, Z-score = 5.02; Wilcoxon Rank Sum Test). When comparing the average expression level of all autosomes and the Z chromosomes, it is clear that the Z chromosome expression in both males and females is outside the autosomal spectrum for all somatic tissues (**Figure 2.2**).

Table 2.1 Comparison of gene expression values in spleen, heart, liver and gonad tissue. Genes are defined as sex-biased if the difference in *log_2* expression between males and females was larger than two-fold and expression was significantly different with (*P* < 0.05 edgeR exactTest, adjusted for multiple testing). Genes are counted as (dosage-) compensated when the difference in *log_2*FC between male and female expression ranged from -0.5 and 0.5.

	Autosomes		**Z-linked**	
	Unbiased	**Sex-biased**	**Compensated**	**Uncompensated**
Spleen	10239 (99.23%)	79 (0.77%)	306 (57.85%)	223 (42.15%)
Heart	9458 (99.79%)	20 (0.21%)	263 (53.35%)	230 (46.65%)
Liver	8900 (99.57%)	38 (0.43%)	252 (56.63%)	193 (43.37%)
Gonad	9694 (85.49%)	1646 (14.51%)	189 (31.29%)	415 (68.71%)

Table 2.2 Comparison between the distributions *log2* expression for male and female autosomal and Z expression in four different tissues. Distributions of autosomal and Z expression are compared using a Wilcoxon Rank Sum Test. Medians are given for the logged expression data.

		Autosomal median	**Z median**	**Difference in median (A-Z)**	***P*-value**	**Z-score**
Male	Spleen	3.72	3.29	0.43	**$3.77x10^{-08}$**	5.50
	Heart	3.10	2.52	0.58	**2.16 x10^{-11}**	6.69
	Liver	2.50	2.02	0.48	**5.08 x10^{-07}**	5.02
	Testes	3.85	3.74	0.11	0.79	0.27
Female	Spleen	3.75	2.80	0.95	**4.46 x10^{-29}**	11.19
	Heart	3.10	2.05	1.05	**3.31 x10^{-29}**	11.22
	Liver	2.53	1.61	0.92	**6.50 x10^{-19}**	8.88
	Ovaries	3.89	3.18	0.71	**3.72 x10^{-20}**	9.20

Significant *P*-values are reported in bold

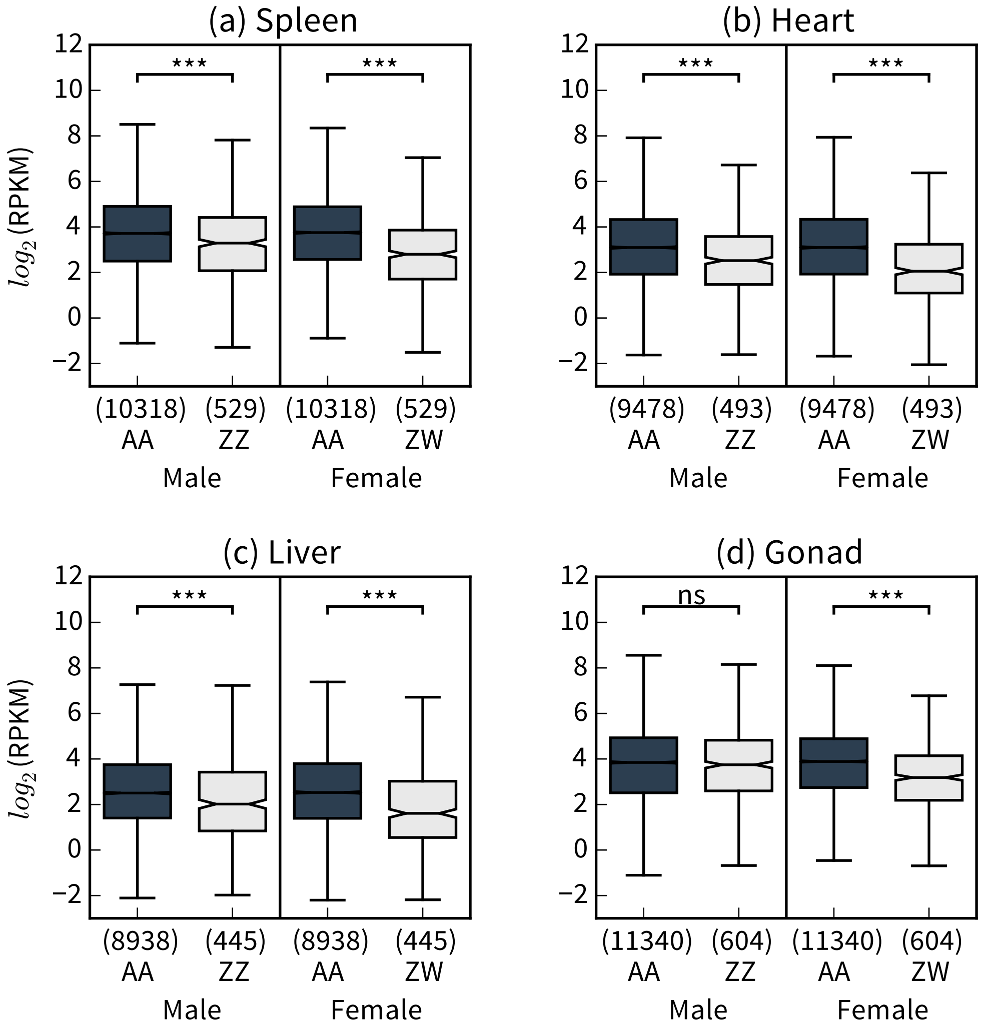


Figure 2.1 Comparison of gene expression measured for autosomal genes (dark grey) and Z-linked genes (light grey) in (a) spleen, (b) heart, (c) liver and (d) gonad tissue in males and females. In all tissues, gene expression for Z-linked genes is significantly lower in comparison to autosomal genes in females. In males, gene expression of Z-linked genes is significantly lower in comparison to autosomal genes in all somatic tissues but not in gonad. Significance levels are indicated as stars (* *P* < 0.05*,* ** *P* < 0.001, *** *P* < 0.0001), differences between distributions were tested using Wilcoxon Rank Sum Tests. The number of genes expressed on the autosomes and Z chromosome(s) are given in brackets for each distribution. Boxes show the interquartile range, notches represent the median of the distribution and whiskers extend to 1.5 times the interquartile range (Q3 + 1.5 x IQR, Q1 – 1.5 x IQR). Outliers are not shown for clarity, but included in all statistical comparisons.

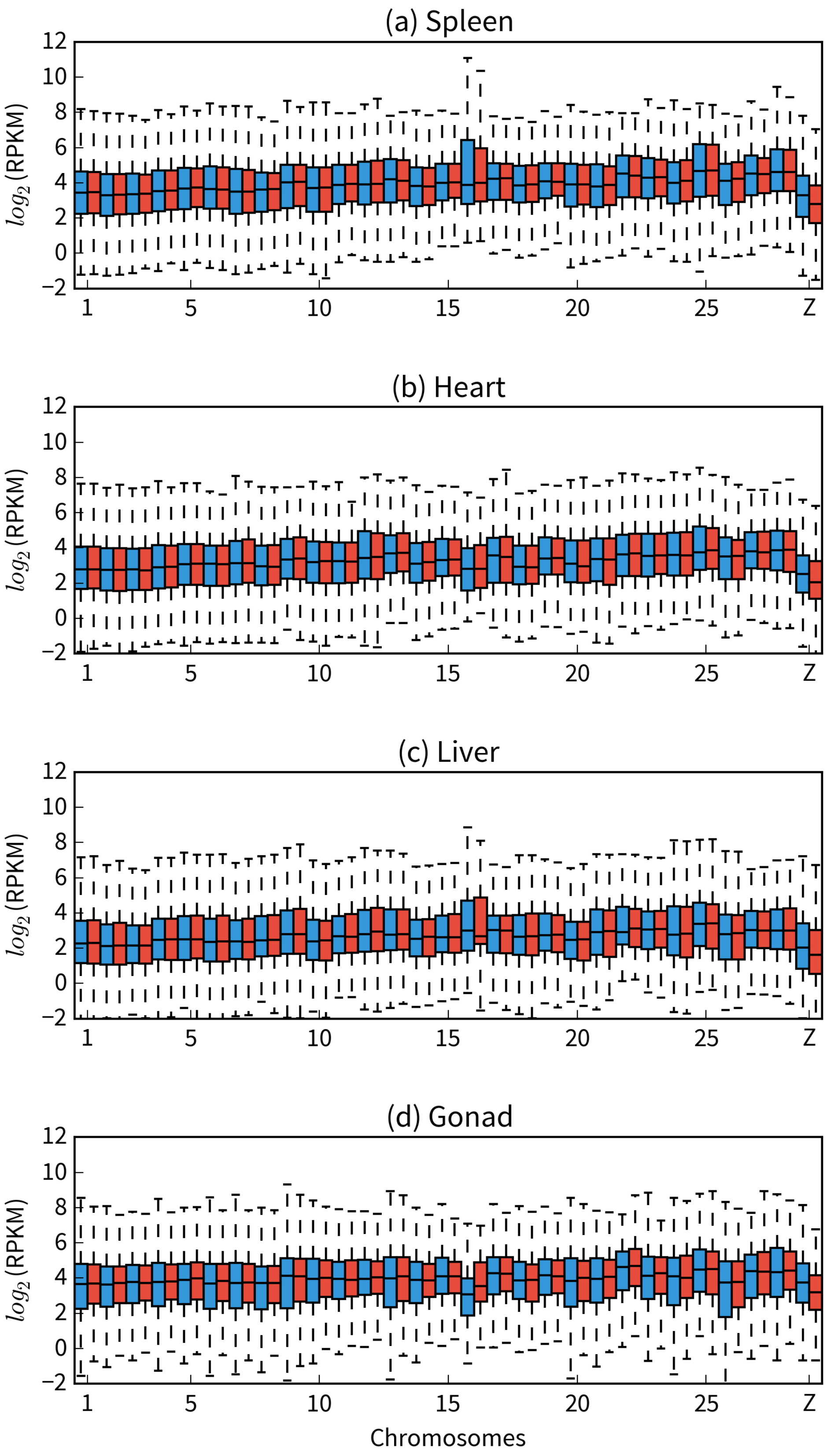


Figure 2.2 Expression level across all chromosomes for males (blue) and females (red) in (a) spleen, (b) heart, (c) liver and (d) gonad tissue.

Z:A ratio comparison

The inclusion of lowly expressed genes may lead to biases when comparing autosomal and Z-linked expression, and previous work illustrated the importance of filtering out lowly expressed and fully inactivated genes (Deng, et al. 2011), which may be non-randomly distributed on sex chromosomes. In order to explore the efficacy of my CPM threshold, I plotted the median Z:A ratios for males (ZZ:AA) and females (Z:AA) across a range of expression thresholds (**Figure 2.3**). Z:A ratios for lowly expressed genes are lower than genes at higher expression levels. More importantly, my results indicate that a CPM threshold > 2 is effective in filtering out lowly expressed genes, as Z:A ratios are similar for higher thresholds. I also found that female Z:A ratios were consistently lower across all threshold levels compared to males, and male Z:A ratios were < 1.0 in all somatic tissues, similar to data from Uebbing, et al. (2015). These lower Z:A ratios in both males and females show considerable tissue-specific variation.

Expression level and dosage compensation

I investigated whether the extent of dosage compensation varies with the magnitude of gene expression, using a quartile-based analysis of Z-linked gene expression (**Figure 2.4**). When grouping the expression of Z-linked genes by male expression, I observed significantly lower female expression in all tissues and quartiles (*P* < 0.05 in all comparisons; Wilcoxon Rank Sum Test), except for the first quartile in gonadic tissue (*P* > 0.05; Wilcoxon Rank Sum Test). In spleen, heart and gonad tissue, the difference between male and female expression is lowest in the first quartile, indicating that dose effects are less prevalent. In liver tissue the lowest difference in gene expression is observed in quartile four. For all other tissues, there is a trend for stronger expression differences in quartiles two to four (*P* < 0.05 in all comparisons; Wilcoxon Rank Sum Test).

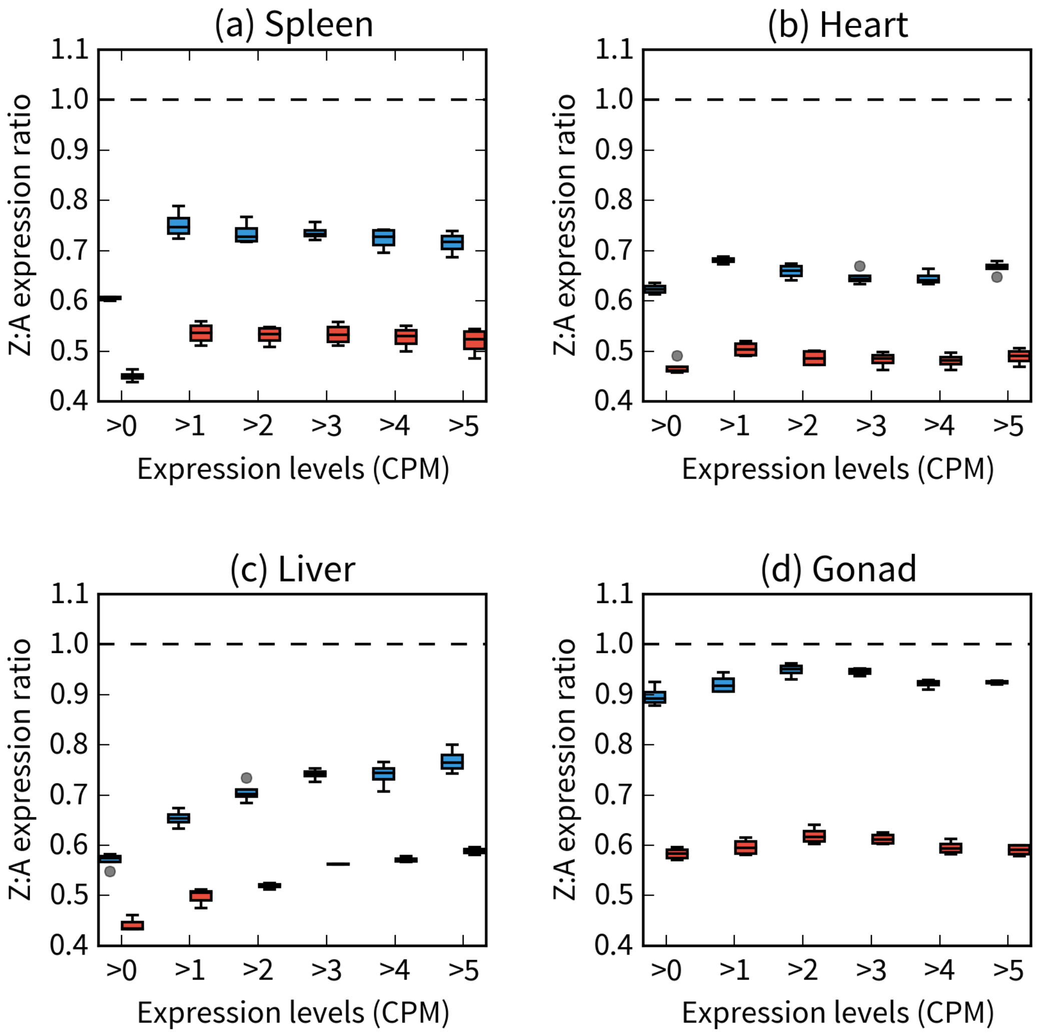


Figure 2.3 Z:A ratio across different CPM thresholds for all male (blue) and all female (red) samples across spleen, heart, liver and gonad tissues. The female Z:A ratio is lower in comparisons to males; however, the male Z:A ratio in spleen, heart and gonad is also lower than the expected ratio of one. Outliers are shown as grey circles.

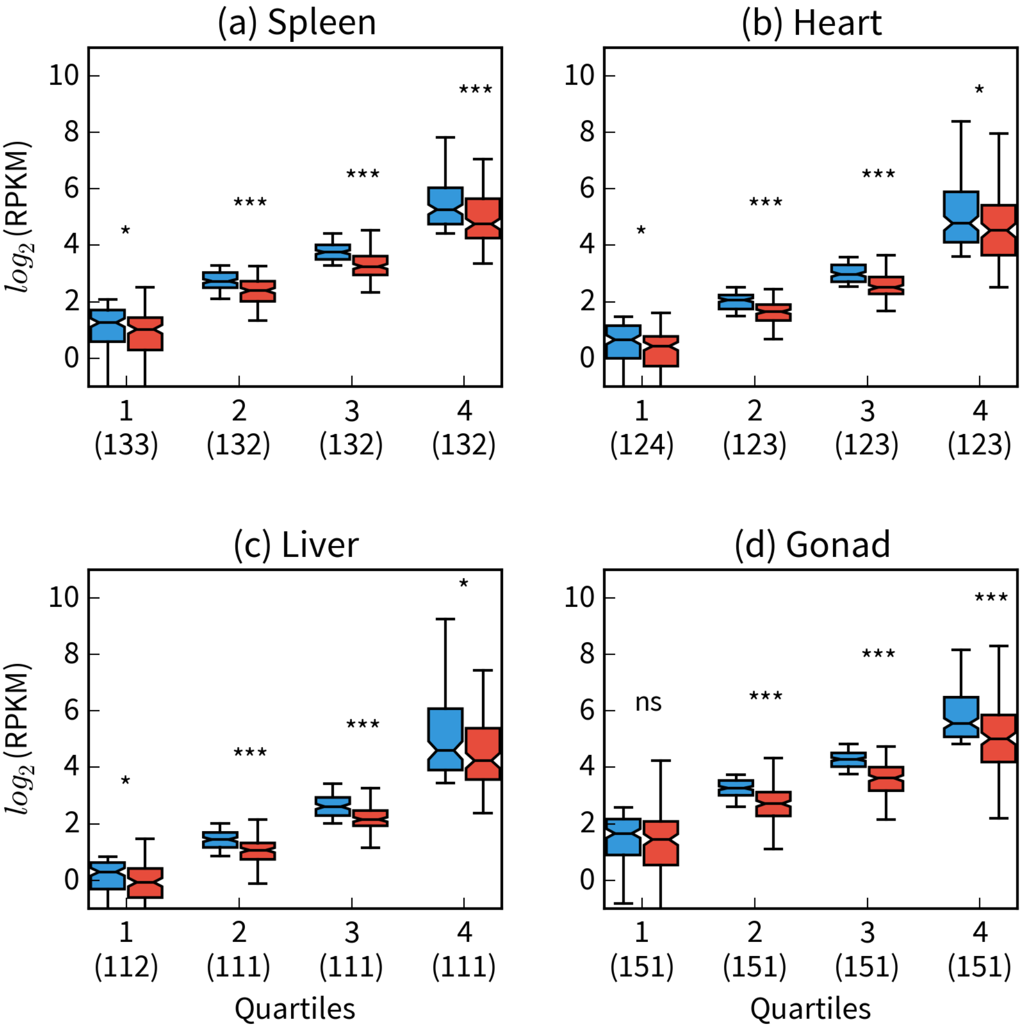


Figure 2.4 *Log₂* transformed CPM expression for genes located on the Z chromosome in spleen, heart, liver and gonad tissue for males (blue) and females (red). Expressed genes are divided into quartiles based on male expression. Significance levels are indicated as stars (* $P < 0.05$, ** $P < 0.001$, *** $P < 0.0001$).

Allele-specific expression and potential for Z chromosome inactivation

The reduction in Z expression in males is also consistent with the possible inactivation of one Z chromosome in males, analogous to the X inactivation observed in therian females (Cooper, et al. 1993; Deakin, et al. 2009). Male Z chromosome inactivation has been suggested by previous work on a limited number of Z-linked loci (Livernois, et al. 2013) and I investigated the potential for Z inactivation using RNA-Seq data. If one copy of the Z chromosome were partially inactivated in males, this would result in SNPs with a significantly greater contribution to the total expression from one allele at heterozygous sites.

As I was primarily interested in the identification of allele-specific expression (ASE) caused by partial Z inactivation, I used a series of stringent filtering criteria in order to reduce the amount ASE that might be the product of *cis*-regulatory variation. After filtering, 35,505 SNPs in spleen, 24,238 SNPs in heart, 18,251 SNPs in liver and 11,783 SNPs in gonad were retained. Of these 826 were detected on the Z chromosome in spleen, 503 in heart, 439 in liver and 405 in gonad tissue. It has been shown previously that the Z chromosome exhibits reduced levels of polymorphism in comparison to the autosomes due to a reduced effective population size (Sundström, et al. 2004), which may explain the low number of detected SNPs on the Z chromosome in comparison to the autosomes.

It is only possible to assess allele-specific expression for Z-linked genes for which at least one valid SNP was detected. Additionally, the detection of allele-specific expression using RNA-Seq depends on a sufficient expression level of an allele and I removed loci from the analysis with < 17 mapped reads. I only called loci as biased if at least 70% of all the reads were from one allele in every sample and the difference was significant in a binomial test ($P < 0.05$). Using these stringent criteria, I found evidence of ASE on the Z chromosome for 10 loci (3.72%) in spleen, 7 (4.09%) in heart, 7 (5.18%) in liver and 13 (10.66%) in gonad; however, only two genes show a consistent signal of ASE across all tissues. I also reduced the ASE threshold to 60%, retaining the same statistical threshold for significance, which resulted in the identification of just one additional locus with ASE in the gonad and one additional locus in the liver. This suggests that the thresholds have not masked a broader pattern of ASE across the Z chromosome.

The dataset contained 10 loci assessed by Livernois, et al. (2013), and only one of these (ENSGAL00000010158; *KANK1*) exhibited significant ASE. There are two potential complications that should be considered when assessing concordance between these approaches. First, of the genes used in the Livernois, et al. (2013) study, five did not contain any valid SNPs in my dataset and I was thus unable to assess ASE in these genes using RNA-Seq. Secondly, the BACs used by Livernois, et al. (2013) were cultured in chicken fibroblasts, and it is possible that the expression patterns in the four tissues used here are different. Additionally, Livernois, et al. (2013) tested for inactivation for any copy of the Z in each cells, where the ASE analysis requires the inactivation of the same Z chromosome. If inactivation were completely random and balanced between the maternal and paternal copy of the Z chromosome across tissues, then ASE would not be detectable. However, although either copy of the X can be inactivated in placental mammals, many genes exhibit a detectable signal of ASE (Payer and Lee 2008; Rozowsky, et al. 2011), and a similar signal in Z inactivation might be expected. In contrast to the analysis of Livernois, et al. (2013), I was able to assess ASE for a much larger number of genes and only recovered a small subset of genes with ASE. These data, therefore, provide little support for a prevalent and widespread inactivation of the Z chromosome.

A different potential explanation for the low number of genes with allele-specific expression could be that large parts of one chromosome are nearly or entirely inactivated and therefore effectively hemizygous in males. If true, this would make the detection polymorphism from RNA-Seq data impossible and I could not assess it in this framework. I also repeated the ASE analysis for the autosomes only and recovered a similar percentage of ASE in comparison to the Z chromosome for all somatic tissues (spleen *P* = 1.0, odds ratio 1.06; heart *P* = 1.0 odds ratio 1.01; liver *P* = 0.50, odds ratio 0.77 Fisher's Exact Test), and a marginally significant difference in gonad (*P* = 0.026, odds ratio

0.49; Fisher's Exact Test). However, given the overall small number of loci affected by ASE, it seems unlikely that this pattern is caused by widespread Z inactivation. Although not the focus of this study, this small subset of genes on the autosomes that show signs of ASE is intriguing. A recent study by Frésard, et al. (2014) indicated that parental imprinting is potentially absent in birds and the fact that the proportion of ASE on the Z chromosome is not significantly higher in comparison to the autosomes further supports the notion that ASE sites on the Z chromosome are not caused by inactivation, but may be the result of *cis*-regulatory variation or other processes.

Ohnologs are preferentially dosage-compensated

If incomplete dosage compensation is sufficient for compensating dosage-sensitive genes, the proportion of dosage-compensated ohnologs on the Z chromosome would be expected to be higher in comparison to non-ohnologs. I tested whether ohnologs are more often dosage-compensated using the expression data and ohnologs obtained from the OhnologsDB (Singh, et al. 2015). The chicken genome contains 5228 (33.71%) annotated ohnologs, of which 223 are annotated on the Z chromosome when using the relaxed set of ohnologs.

In order to determine whether ohnologs are preferentially dosage-compensated, I first compared the log_2 fold change between female and male expression for Z-linked ohnologs and non-ohnologs (**Figure 2.5**). The difference in expression between females and males (log_2FC) was significantly lower for ohnologs than non-ohnologs (spleen $P < 0.0001$, Z-score = 5.95; heart $P < 0.0001$, Z-score = 4.57; liver $P < 0.0001$, Z-score = 5.22; gonad $P < 0.0001$, Z-score = 4.89; Wilcoxon Rank-Sum Test), suggesting a higher degree of dosage compensation. Additionally, the proportion of dosage-compensated ohnologs (log_2FC range from -0.5 to 0.5) was significantly higher in comparison to non-ohnologs in all tissues (P-value < 0.0001 in all comparisons; Fisher's Exact Test, **Table 2.3**). This is also the case when I used a wider range of log_2FC (-0.6 to 0.6), similar to the mean expression change for female one-dose genes reported by Malone, et al. (2012) (**Table 2.4**). Additionally, I used the strict set of ohnologs from the OhnologsDB, with 2489 ohnologs annotated in the chicken genome and 106 on the Z chromosome, recovering similar results (**Table 2.5**, **Figure 2.6**).

An alternative explanation for the high degree of dosage compensation among ohnologs is that all paralogs, even those that originate in single-gene duplications, are dosage-compensated. I tested this hypothesis by extracting Z-linked paralogs from the Ensembl database (Cunningham, et al. 2015) that originated in single-gene duplication events. These paralogs do not show a higher proportion of dosage compensation ($P > 0.05$ in all comparisons; Fisher's Exact Test; **Table 2.6**) compared to all other genes on the Z chromosome. This indicates that the higher degree of dosage compensation among ohnologs is not a property of paralogs in general, and that the mode of duplication has an important impact on the evolution of gene-by-gene dosage compensation.

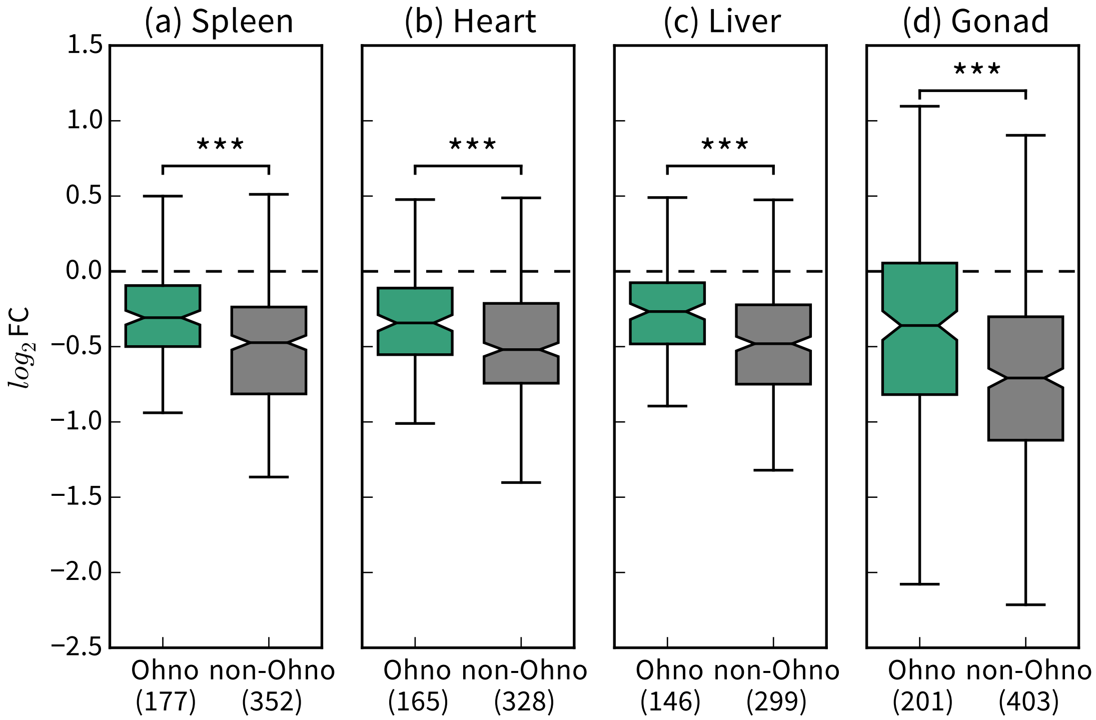


Figure 2.5 Comparison of log_2-transformed fold-change between female and male expression for ohnologs (green) and non-ohnologs (grey) on the Z chromosome in (a) spleen, (b) heart, (c) liver and (d) gonad. The number of genes in the distributions is given in brackets. Negative fold-changes indicate higher male expression; positive fold-changes indicate stronger female expression. Significance levels are indicated as stars (* $P < 0.05$, ** $P < 0.001$, *** $P < 0.0001$), differences between distributions were tested using Wilcoxon Rank Sum Tests. Outliers are not shown for clarity, but included in all statistical comparisons.

Table 2.3 Contingency tables for all four tissues, comparing the proportion of dosage-compensated (DC) and uncompensated (U) ohnologs to non-ohnologs using a Fisher's Exact Test. Genes are called as dosage-compensated if the *log*$_2$FC ranges from -0.5 to 0.5.

	Ohnolog		Non-ohnolog			
	DC	**U**	**DC**	**U**	***P*-value**	**Odds ratio**
Spleen	126 (71.19%)	51 (28.81%)	180 (51.14%)	172 (48.86%)	**1.08 x 10^{-5}**	2.36
Heart	111 (67.27%)	54 (32.73%)	152 (46.34%)	176 (53.66%)	**1.06 x 10^{-5}**	2.38
Liver	105 (71.92%)	41 (28.08%)	147 (49.16%)	152 (50.84%)	**6.52 x 10^{-6}**	2.65
Gonad	86 (42.79%)	115 (57.21%)	103 (25.56%)	300 (74.44%)	**2.57 x 10^{-5}**	2.18

Significant *P*-values are reported in bold

Table 2.4 Contingency tables for all four tissues, comparing the proportion of dosage-compensated (DC) and uncompensated (U) ohnologs to non-ohnologs using a Fisher's Exact Test. Genes are called as dosage-compensated if the *log*$_2$FC ranges from -0.6 to 0.6.

	Ohnolog		Non-ohnolog			
	DC	**U**	**DC**	**U**	***P*-value**	**Odds ratio**
Spleen	149 (84.18%)	28 (15.82%)	200 (56.82%)	152 (43.18%)	**1.52 x10^{-10}**	4.04
Heart	129(78.18%)	36 (21.82%)	183 (55.79%)	145 (44.21%)	**1.03 x10^{-6}**	2.84
Liver	116 (79.45%)	30 (20.55%)	175 (58.53%)	124(41.47%)	**1.20 x10^{-5}**	2.74
Gonad	105 (52.24%)	96 (47.76%)	135 (33.50%)	268 (66.50%)	**1.41 x 10^{-5}**	2.17

Significant *P*-values are reported in bold

Table 2.5 Contingency tables for all four tissues, comparing the proportion of dosage-compensated (DC) and uncompensated (U) ohnologs to non-ohnologs using a Fisher's Exact Test. Ohnologs are from the strict subset of the OhnologsDB.

	Ohnolog		**Non-ohnolog**			
	DC	**U**	**DC**	**U**	***P*-value**	**Odds ratio**
Spleen	63 (75.12%)	22 (25.88%)	243 (54.73%)	201 (45.27%)	**0.0011**	2.37
Heart	61 (72.62%)	23 (27.38%)	202 (49.39%)	207 (50.61%)	**0.0001**	2.72
Liver	53 (68.83%)	24 (31.17%)	199 (54.08%)	169 (45.92%)	**0.0224**	1.87
Gonad	47 (47.96%)	51 (52.04%)	142 (28.06%)	364 (71.94%)	**0.0002**	2.36

Significant *P*-values are reported in bold

Table 2.6 Contingency tables for all four tissues, comparing the proportion of dosage-compensated and uncompensated paralogs (excluding ohnologs) to all other genes (including ohnologs) using a Fisher's Exact Test.

	Paralogs (excluding ohnologs)		**Other**			
	DC	**U**	**DC**	**U**	***P*-value**	**Odds ratio**
Spleen	96 (61.15%)	61 (38.85%)	210 (56.45%)	162 (43.55%)	0.34	1.21
Heart	76 (51.70%)	71 (48.30%)	187 (54.05%)	159 (45.95%)	0.69	0.91
Liver	76 (57.14%)	57 (42.86%)	176 (56.41%)	136 (43.59%)	0.92	1.03
Gonad	55 (28.06%)	141 (71.94%)	134 (32.84%)	274 (67.16%)	0.26	0.80

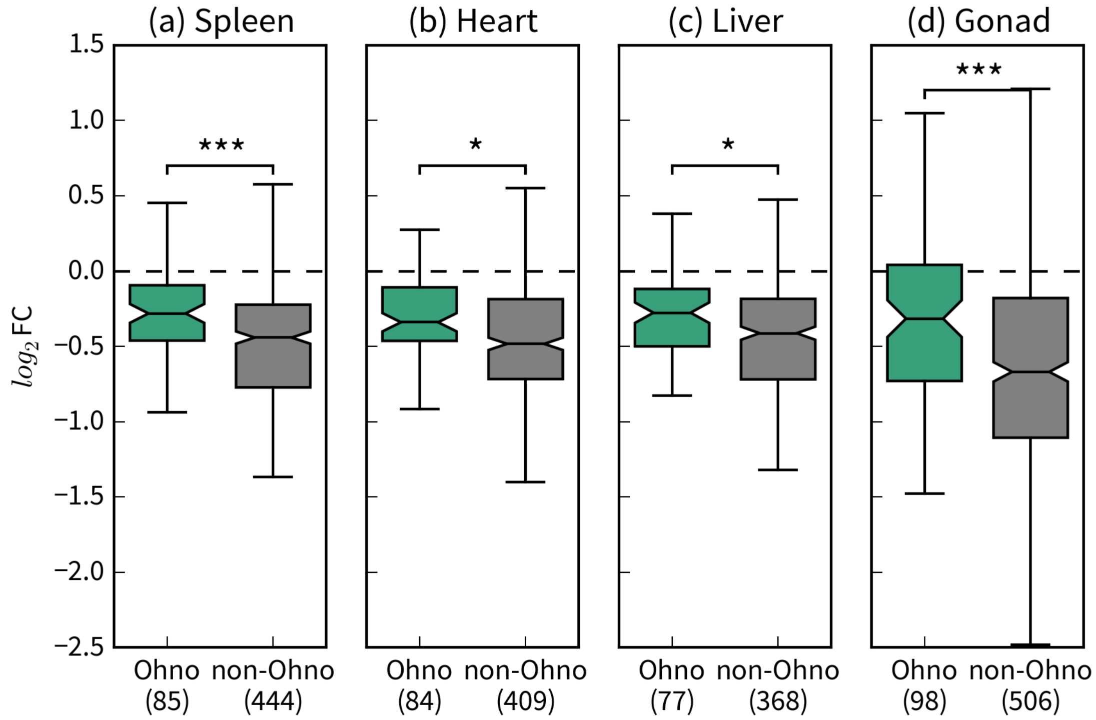


Figure 2.6 Comparison of *log*$_2$-transformed fold-change between female and male expression for ohnologs (green), taken from the strict OhnologsDB subset, and non-ohnologs (grey) on the Z chromosome in (a) spleen, (b) heart, (c) liver and (d) gonad. The number of genes in the distributions is given in brackets. Negative fold-changes indicate higher male expression; positive fold-changes indicate stronger female expression. Significance levels are indicated as stars (* $P < 0.05$, ** $P < 0.001$, *** $P < 0.0001$), differences between distributions were tested using Wilcoxon Rank Sum Tests. Outliers are not shown for clarity, but included in all statistical comparisons.

Older Z chromosome parts contain fewer ohnologs

Sex chromosome divergence can drive the movement of some gene classes off the sex chromosomes (Emerson, et al. 2004; Potrzebowski, et al. 2008; Vibranovski, et al. 2009) and an out of Z migration for dosage-sensitive genes might be expected. Overall, the proportion of ohnologs is not significantly different between the Z (764 coding genes) and the genomic background (14744 coding genes) (*P* = 0.19, odds ratio = 0.89; Fisher's Exact Test), suggesting that the Z chromosome is not depleted of ohnologs and that dosage-sensitive gene have not moved off the Z. However, the Z chromosome contains at least four strata, where recombination was suppressed between the Z and W at different times, spanning roughly 130 million years (Wright, et al. 2012). I divided the chromosome into an old and young part along the border of stratum 3, resulting in two almost equally sized regions of the Z chromosome. Given 223 ohnologs located on the Z chromosome, I expected that half of these would be located in the old and half in the young part of the chromosome. However, the number of ohnologs in the older half of the chromosome is significantly less than expected (X^2=22.605, *P* < 0.0001; Chi-Square Test) and also significantly less when accounting for the difference in gene content (*P* < 0.05, odds ratio = 0.62; Fisher's Exact Test). This could indicate that some ohnologs may have relocated during the early evolution of the Z chromosome. When I compared the proportion of dosage-compensated ohnologs between old and young parts of the Z chromosome, I did not detect a significantly higher proportion of dosage-compensated ohnologs in older parts (*P* > 0.05 in all comparisons; Fisher's Exact Test), suggesting that dosage compensation of ohnologs occurs relatively quickly following W chromosome gene loss. Alternatively, this bias could be an artefact of the ancestral ohnolog distribution, as the WGD events precede the formation of the sex chromosome system.

Dosage compensation of ohnologs across tissues

The degree of dosage compensation is similar in all somatic tissues (*P* > 0.05 in all comparisons; Fisher's Exact Test; **Table 2.7**), and greater in the soma compared to the gonad (*P* < 0.0001 in all comparisons; Fisher's Exact Test; **Table 2.7**). Tissues can be seen as a form of functional compartmentalization, and the same gene can show a diverse range of expression patterns in different tissues. For this reason, similar overall dosage compensation could hide an underlying pattern of pleiotropic expression. Dosage sensitivity may in fact be tissue dependent and can result in gene-by-gene dosage compensation (Mank and Ellegren 2008).

I also investigated the overlap of dosage-compensated ohnologs across tissues. A set of 68 of 223 ohnologs was dosage-compensated in all somatic tissues; however, I detected substantial variation (**Figure 2.7**). Of the 68 ohnologs that are dosage-compensated in all somatic tissues, only 36 are also dosage-compensated in gonad, showing that only a small core set of ohnologs are dosage-sensitive across all tissues. In gonad, a unique set of 50 ohnologs was dosage-compensated. In combination with the overall lower degree of dosage compensation in gonad, this suggests different dosage compensation patterns in comparison to the somatic tissues.

Table 2.7 Comparison of the degree of dosage compensation between all four tissues using a Fisher's Exact Test. Cells show *P*-values, numbers in brackets are the odds ratio. Data for the contingency tables is shown in **Table 2.1**. Auto-comparisons and mirrors are not calculated.

	Spleen	**Heart**	**Liver**	**Gonad**
Spleen				
Heart	0.17 (0.83)			
Liver	0.74 (0.95)	0.32 (1.14)		
Gonad	**2.28 x10^{-19}** (0.33)	**1.73 x10^{-13}** (0.40)	**2.25 x10^{-16}** (0.35)	

Significant *P*-values are reported in bold

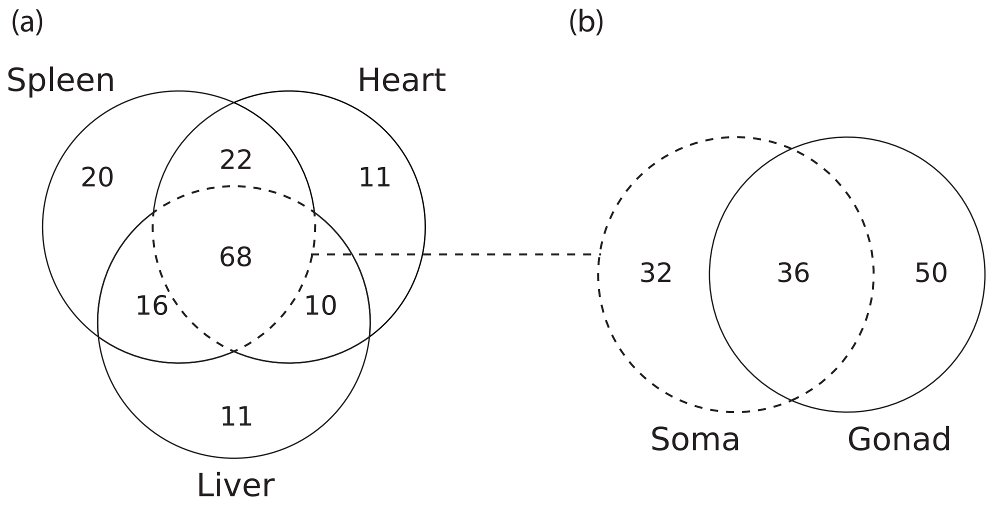


Figure 2.7 (a) Overlap between dosage-compensated ohnologs in the three somatic tissues. (b) Overlap between dosage-compensated genes in the soma (spleen, heart and liver) and gonad tissue. Circles represent the total of dosage-compensated ohnologs in a tissue and numbers indicate the overlap between sets.

Functional annotation of Z-linked ohnologs

Ohnologs are associated with a wide range of functions, such as signalling pathways (Blomme, et al. 2006) and protein complexes (Makino, et al. 2009). This functional enrichment is the reason for the hypothesized dosage sensitivity and may explain why ohnologs are preferentially dosage-compensated. I used the G:profiler online tool (Reimand, et al. 2011) to test for functional overrepresentation of gene ontology (GO) terms (Ashburner, et al. 2000) and compared the 223 ohnologs on the Z chromosome to the genomic background. I detected 13 significantly enriched GO terms (P < 0.05 after correction for multiple testing), many of which are associated with membrane proteins, cell locomotion, and localization (**Table 2.8**). I also detected enrichment for the oncostatin-M-mediated signalling pathway, a cytokine that may be important in cell proliferation and multiple diseases, such as cancer (Dey, et al. 2013). Additionally, I used the CORUM (Ruepp, et al. 2010) database to test for the overrepresentation of protein complexes among ohnologs, but did not detect a significant enrichment in the 223 ohnologs on the Z chromosome in comparison to other Z-linked genes (P > 0.05, Fisher's Exact Test).

Table 2.8 Significantly overrepresented GO terms (*P* <0.05 after g:SCS multiple testing correction) for the 223 ohnologs located on the chicken Z chromosome. Overrepresentation was tested using the 223 ohnologs as a target set and the genomic background (autosomal genes and other non-ohnologs on the Z chromosome). Only genes that have annotated GO terms were considered in the analysis. GO term ID is prefaced by the functional category: BP (biological process), CC (cellular component) and MF (molecular function). The number of genes in the chicken genome annotated with a specific term (T) is shown in column three, and the overlap between the query (Q) set and the annotated background (T) is shown in column four (Q&T).

GO term ID		***P*-value**	**T**	**Q&T**	**GO term description**
BP	GO:0038165	**0.0098**	3	3	oncostatin-M-mediated signalling pathway
BP	GO:0040011	**0.0135**	878	33	locomotion
BP	GO:0051179	**0.0011**	3523	91	localization
BP	GO:0051674	**0.0004**	746	33	localization of cell
BP	GO:1902578	**0.0028**	2448	69	single-organism localization
BP	GO:0044765	**0.0033**	2308	66	single-organism transport
BP	GO:0006928	**0.0092**	986	36	movement of cell or subcellular component
BP	GO:0048870	**0.0011**	745	32	cell motility
BP	GO:0016477	**0.0028**	701	30	cell migration
CC	GO:0016020	**0.0166**	4932	112	membrane
CC	GO:0098589	**0.0003**	457	25	membrane region
CC	GO:0044459	**0.0301**	1259	41	plasma membrane part
CC	GO:0098590	**0.0074**	373	20	plasma membrane region
MF	GO:0004924	**0.0098**	3	3	oncostatin-M receptor activity

Significant *P*-values are shown in bold

Discussion

My analyses of dosage compensation and ohnologs on the chicken Z chromosome provide novel insights into the nature of incomplete dosage compensation. I confirm previous reports of incomplete dosage compensation in chicken (Ellegren, et al. 2007; Itoh, et al. 2007; Uebbing, et al. 2015) and show that ohnologs are preferentially dosage-compensated on the chicken Z chromosome, indicating that incomplete dosage compensation can effectively balance dosage-sensitive genes. Even though the average expression of the Z chromosome is consistently lower in females as a function of incomplete dosage compensation, a considerable number of Z-linked genes show equal expression between males and females. Moreover, selection for compensation of dosage-sensitive genes appears to act relatively quickly, as there is no significant difference in the proportion of dosage-compensated ohnologs in younger regions of the avian Z chromosome compared to older regions.

The X chromosomal expression in mammals is reduced compared to the autosomes, potentially as a consequence of X inactivation (Julien, et al. 2012; Xiong, et al. 2010). It has been suggested that selection for the compensation of dosage-sensitive genes could have driven the evolution of X inactivation in therian mammals. Similarly, I also observed a reduction in Z expression in somatic tissues in males (Itoh, et al. 2007). The reduced expression of the Z chromosome compared to the autosomes in males is not as pronounced as in females (**Table 2.2, Figure 2.1, Figure 2.2**) and there are several possible explanations for this pattern. The reduction has been suggested to result through Z inactivation that affects only some parts of the chromosome (Graves 2014; Livernois, et al. 2013). However, my assessment of ASE suggests that inactivation is not a major mechanism affecting Z chromosome expression in males. An alternative explanation for the lower Z expression may be that the ancestral expression level of the Z chromosome, before the differentiation of the sex chromosomes, was already on average on the lower end of the expression spectrum (Brawand, et al. 2011; Julien, et al. 2012). Finally, it is possible that dosage sensitive genes have moved off the Z, as the mammalian

X chromosome is depleted of genes requiring high transcription rates as a result of haploid expression in females (Hurst, et al. 2015). My analysis suggests that although there is some potential for movement of dosage-sensitive genes off the Z chromosome, the effect is confined to the oldest regions of the Z chromosome and is not substantial enough to explain the reduced expression in males.

Chapter 3

Inferring paralogs and expression divergence across multiple bird species using RNA-Seq data

Summary

Single gene duplications are an important source of novel genetic material. After gene duplication, pairs of gene copies (paralogs) diverge either through neutral or selective forces. This functional divergence can be brought about through changes in protein-coding sequence or regulatory mutations. How paralogs evolve after the initial duplication event has been intensely studied, with multiple models describing different evolutionary outcomes. In order to test the predictions made by these models, comparative genomic data is needed. These data can be used to reconstruct the evolutionary history of orthologs (genes in different species with a common ancestor) and paralogs. However, genomic data is not available in many ‘non-model’ species, limiting these analyses to well-studied clades. In contrast, the availability of *de novo* RNA-Seq assemblies has facilitated the analyses of transcriptomes in ‘non-model’ species. If comparative RNA-Seq could be used for paralog detection, it would not only allow for comparative analyses in many 'non-model' species but would also enable investigations of paralog gene expression divergence in a phylogenetic framework. In this chapter, I investigated the suitability of *de novo* RNA-Seq assemblies for paralog detection and comparative gene expression analyses using a combination of established genomic tools and novel filtering strategies. These analyses highlight current limitations and pitfalls of this approach and suggest that *de novo* RNA-Seq assemblies may be unsuitable for paralog detection.

Introduction

Single gene duplications are a major source of genetic variation and play an important role in the evolution of novel functions (Zhang 2003). Gene duplication creates two gene copies (paralogs), increasing the availability of genetic material for evolution to act on (Ohno 1970; reviewed by Hahn 2009; Innan and Kondrashov 2010). Gene duplication is followed by functional divergence of paralogs, either through neutral processes - most often resulting in pseudogenisation - or as a result of selection (see chapter 1). The term 'functional divergence' is ambiguous but can be mainly viewed as changes in protein-coding DNA sequence (Zhang 2003), regulatory changes that affect gene expression (Force, et al. 1999), or post-translational modifications (Nguyen Ba, et al. 2014). The availability of genetic data in combination with models of sequence evolution implemented, for example, by PAML (Yang 2007), can be used to test hypotheses regarding the evolution of paralogs.

Previous studies of paralog divergence have focused mainly on changes in protein-coding DNA sequence. Before the advent of high-throughput sequencing, coding regions of paralogs were amplified via PCR (using specific primers) and sequenced using Sanger sequencing. For example, the duplication of a pancreatic ribonuclease gene in a colobine monkey (*RNASE1a* and *RNASE1b*) was investigated using this technique (Zhang, et al. 2002). In this case, several adaptive nonsynonymous substitutions in the *RNASE1b* gene occurred after duplication, shifting *RNASE1b* function towards enzymatic activity in a new chemical environment (Zhang 2003; Zhang, et al. 2002). This duplication has occurred independently in African and Asian clades of leaf-eating monkeys with similar patterns of functional divergence, constituting a case of parallel, adaptive evolution (Zhang 2006).

The emergence of high-throughput sequencing (e.g., Illumina HiSeq) has led to the generation of draft genomes for many species, particularly in well-studied clades. With increasing data availability and quality, comparative genomics has been used to identify adaptive, species-specific duplications. For example, the availability of the human reference genome in combination with the resequencing of specific regions on the human chromosome 1 revealed that the Slit-Robo Rho GTPase activating protein 2 (*SRGAP2*) gene underwent an incomplete duplication event in the human lineage ca. 3.4 million years ago (Dennis, et al. 2012). This duplication was followed by two larger segmental duplications that resulted in four paralogs, two of which (*SRGAP2B* and *SRGAP2D*) are likely to be pseudogenised and non-functional, whereas *SRGAP2A* and *SRGAP2C* are actively expressed and interact competitively, resulting in a derived developmental effect (Dennis, et al. 2012).

These comparatively recent duplications illustrate how, over large phylogenetic distances, duplication and loss of genes can create a complex landscape of gene families that can be investigated by comparing the draft genomes across clades of interest (Fortna, et al. 2004; Hahn, et al. 2007). Many databases, such as Ensembl (Cunningham, et al. 2015), InParanoid (Sonnhammer and Östlund 2015) or OMA (Altenhoff, et al. 2015), offer pre-computed groups of orthologs and paralogs for a wide range of species with available draft genomes. These databases usually use a combination of all-vs-all protein sequence alignments, often performed with BLAST (Altschul, et al. 1990) or Smith-Waterman algorithms (Smith and Waterman 1981), and different clustering techniques, such as Markov Clustering (Enright, et al. 2002), to detect orthologs and paralogs. These tools facilitate the detailed study of gene family evolution and constitute valuable resources for investigating the functional divergence of paralogs through changes in protein-coding DNA sequence.

Despite the increasing availability of draft genomes, many organisms studied to address a range of biological questions still lack high-quality genomic resources. In the absence of a reference genome, RNA-Seq experiments have become a popular and relatively inexpensive route to characterising the gene content and expression levels in non-model organisms (Wang, et al. 2009). RNA-Seq can be performed without sequenced genomes but relies on the *de novo* assembly of the transcriptome. *De novo* assembly of transcriptomes using RNA-Seq reads is a computationally challenging problem, similar to the *de novo* assembly of genomes, but does not require the physical mapping of genes onto chromosomes. Specialised assemblers, such as Trinity (Grabherr, et al. 2011; Haas, et al. 2013), AbySS (Birol, et al. 2009) and SOAPdenovo-Trans (Xie, et al. 2014), have been developed to allow the efficient reconstruction of a *de novo* transcriptome.

If *de novo* RNA-Seq data were sufficient for the identification of paralogous genes, it would be possible to vastly extend analyses of gene family evolution to species without high quality reference genomes. Additionally, it would be possible to combine RNA-Seq data with available genomic data, which may aid the reconstruction of gain loss and duplication of genes. RNA-Seq data also has the added benefit of providing gene expression estimates for paralogs, which would facilitate studies of paralog gene expression divergence in a phylogenetic framework. Several lines of evidence suggest that functional divergence of paralogs occurs through regulatory changes. For example, studies of human paralog pairs indicate that differential gene expression between paralogs (Makova and Li 2003) evolves quickly, a pattern that has also been reported in a comparative study of species-specific paralogs in human and mouse (Huminiecki and Wolfe 2004). Following gene duplication, paralogs evolve narrower, often more tissue-specific, expression patterns (Huerta-Cepas, et al. 2011; but see Schmitz, et al. 2016). Consequently, large gene families, created by multiple duplication and loss events, often contain paralogs with high tissue-specificity (Huminiecki and Wolfe 2004). Within species, paralogs also show more divergent expression patterns than orthologs do between species, which supports the hypothesis that paralogs diverge quickly (Chen and Zhang 2012; Rogozin, et al. 2014).

Existing studies of comparative gene expression either rely on the availability of draft genomes (e.g. Brawand, et al. 2011; Coolon, et al. 2014; Rhind, et al. 2011) or when using *de novo* RNA-Seq assemblies only investigate a core set of one-to-one orthologous genes compared to a well-annotated genome of a closely related species (e.g.Harrison, et al. 2015). This approach ignores the potentially important role of gene duplication in the evolution of species differences. Identifying paralogous genes from RNA-Seq assemblies could provide a route to address this limitation; however, it is still unclear how well suited *de novo* RNA-Seq data is for the reconstruction of gene family evolution.

In this chapter, I assess the potential of *de novo* RNA-Seq assemblies for the reconstruction of gene family histories and the identification of lineage-specific paralogs. *De novo* RNA-Seq assemblies generate a large number of potential transcripts, many of which may not be biologically relevant, and I investigate if this negatively affects the reconstruction of gene family histories. I explore whether RNA-Seq and DNA-Seq data can be combined to improve the power of paralog detection, and I test the reliability of the results by comparing estimated duplication rates to previous genomic estimates. I also assess whether paralogs inferred from RNA-Seq data can be validated using genomic data.

To do this, I used a comparative RNA-Seq dataset of six Galloanserae bird species and constructed *de novo* transcriptome assemblies. I then used these data, in combination with all available bird genomes from the Ensembl database (Cunningham, et al. 2015), to reconstruct gene family histories for both the RNA-Seq based assemblies, the DNA-Seq based assemblies and a combination of all data with the Orthologous MAtrix (OMA; Altenhoff, et al. 2015) tool. The influence of *ab initio* protein predictions and noise in the RNA-Seq data was assessed by comparing reconstruction results between RNA-Seq and DNA-Seq data, with and without a series of filters designed to remove potential false-positives. These analyses highlight the challenges, limitations and potential pitfalls of using RNA-Seq data for the reconstruction of gene family evolution and paralogs in particular.

Material and Methods

In order to reconstruct gene family evolution and detect paralogs from RNA-Seq data, I designed a bioinformatics pipeline that incorporates both *de novo* RNA-Seq and DNA-Seq data (**Figure 3.1**). The pipeline builds *de novo* RNA-Seq assemblies from six different bird species (labelled ‘De novo RNA-Seq’, **Figure 3.1**) and infers likely protein-coding sequences. For DNA-Seq data, precomputed protein sequences were used (labelled ‘DNA-Seq’, **Figure 3.1**). I ran the OMA analysis in three iterations (blue boxes, **Figure 3.1**) to compare results between single data types, and when different data types were combined. I first ran OMA separately for RNA-Seq and DNA-Seq data (boxes ‘OMA run 1’ and ‘OMA run 2’, **Figure 3.1**), which enabled the detection of patterns that are specific to the RNA-Seq and the DNA-Seq data. I subsequently ran OMA using combined gene model predictions from both RNA and DNA data (box ‘OMA run 3’, **Figure 3.1**). If the results from RNA-Seq and DNA-Seq based assemblies were comparable, this approach should increase the power to reconstruct gene family evolution and provide better phylogenetic coverage. Using these three runs, I assessed consistency by doing i) direct comparisons between RNA- and DNA-Seq data in the combined run; ii) comparing the amount of loss and gains between DNA-Seq and RNA-Seq data; iii) reconstructing ancestral gene family composition in the combined run. In the following paragraphs, I provide detailed descriptions of all steps implemented in the bioinformatics pipeline.

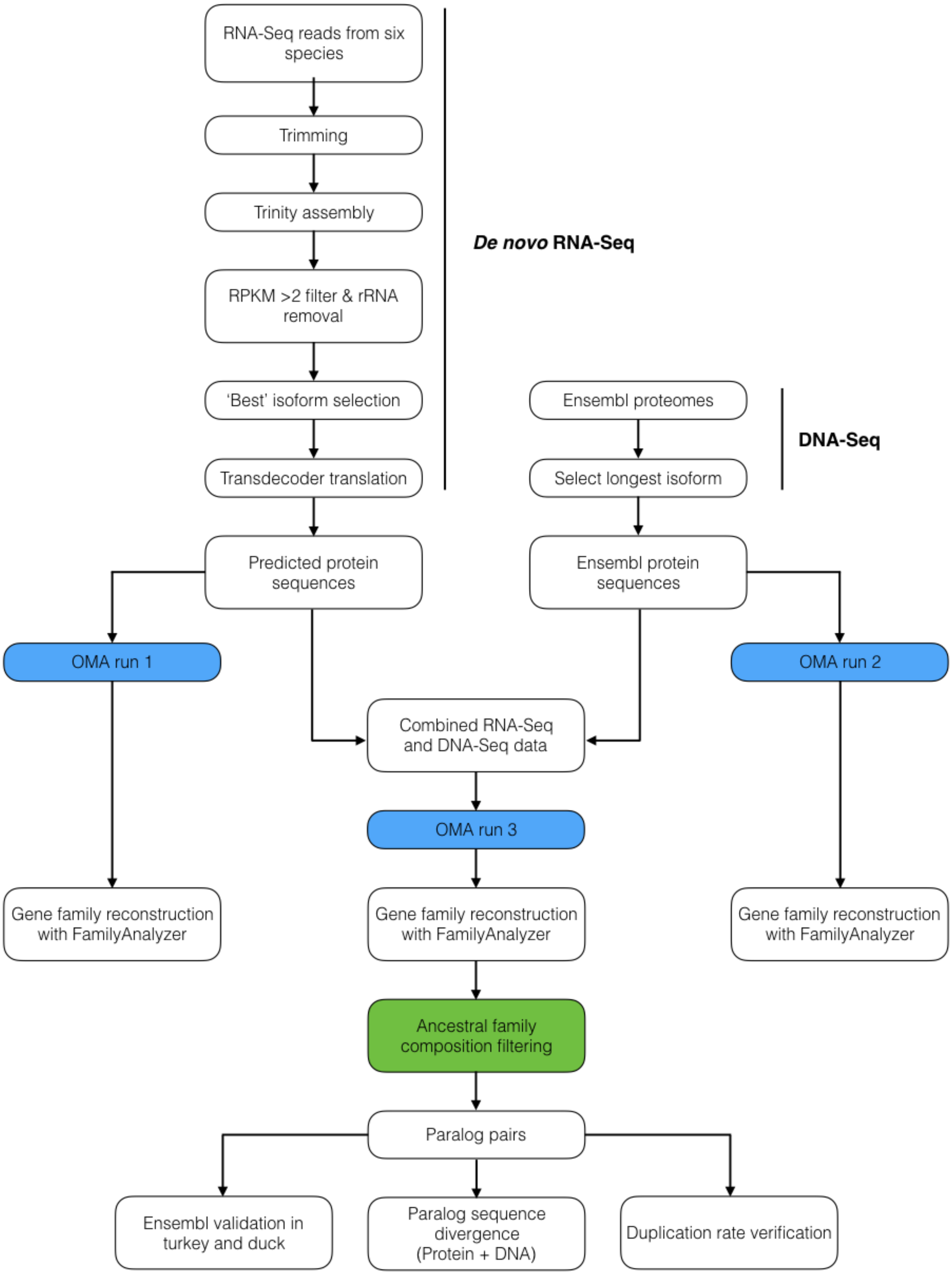

RNA-Seq reads from six species
Trimming
Trinity assembly
RPKM >2 filter & rRNA removal
'Best' isoform selection
Transdecoder translation
De novo RNA-Seq
Ensembl proteomes
Select longest isoform
DNA-Seq
Predicted protein sequences
Ensembl protein sequences
OMA run 1
OMA run 2
Combined RNA-Seq and DNA-Seq data
OMA run 3
Gene family reconstruction with FamilyAnalyzer
Gene family reconstruction with FamilyAnalyzer
Gene family reconstruction with FamilyAnalyzer
Ancestral family composition filtering
Paralog pairs
Ensembl validation in turkey and duck
Paralog sequence divergence (Protein + DNA)
Duplication rate verification

Figure 3.1 Design of the bioinformatics pipeline used to reconstruct gene family evolution, detect paralog pairs and evaluate differences between DNA-Seq and RNA-Seq based analyses. OMA runs are coloured in blue; the ancestral family composition filtering is shown in green and was implemented in a patched version of FamilyAnalyzer. Arrows indicate the flow of data; boxes represent computational steps.

RNA-Seq assembly and gene expression calculation

In order to assess if RNA-Seq data can be used for paralog detection, I obtained RNA-Seq libraries from six species of Galloanserae (swan goose, *Anser cygnoides*; mallard duck, *Anas platyrhynchos*; wild turkey, *Meleagris gallopavo*; helmeted guineafowl, *Numida meleagris*; Indian peafowl, *Pavo cristatus* and common pheasant, *Phasianus colchicus*) stored in the Short Read Archive under Bio Project ID PRJNA271731 (box ‘RNA-Seq reads from six species’, **Figure 3.1**). Full details of the sample preparation are described in (Harrison, et al. 2015). Briefly, the left gonad and spleen from five adult males and five adult females were sequenced for all species except for pheasant and turkey. In pheasant, six male and five female libraries were sequenced for both spleen and gonad; in turkey, only four male and two female spleen samples were sequenced. All libraries were sequenced as 100bp paired-end reads using Illumina HiSeq 2000.

The data processing and assembly methods presented in this paragraph were aided by scripts and previous work done by Dr Peter Harrison.
All libraries were quality inspected using FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). To ensure high input read quality, Trimmomatic v.22 (Bolger, et al. 2014) was used to trim reads (LEADING:5, TRAILING:5 SLIDINGWINDOW4:5, MINLEN:25); Illumina adapter sequences were also removed (box ‘Trimming’, **Figure 3.1**). All samples for one species were subsequently concatenated and used as input for the Trinity v2.0.2 *de novo* assembler with default settings and enabled *in silico* normalisation (Grabherr, et al. 2011; Haas, et al. 2013) (box ‘Trinity assembly’, **Figure 3.1**). The six *de novo* assemblies were then used as a species reference for all subsequent analyses. An overview of the assembly statistics is shown in **Table 3.1**.

Table 3.1 Trinity assembly statistics for all six bird species. Statistics are based on all transcript contigs and not on the longest isoform per predicted gene model. N50 represents the contig length at which half of all contigs in the assembly are longer than this number (Miller, et al. 2010)

Species	Total assembled bases	Total trinity 'genes'	Total trinity transcripts	Percent GC	Median contig length	Average contig length	N50
A. cygnoides	1,107,193,855	845,632	1,055,924	45,80%	455	1,048,55	2,292
A. platyrhynchos	922,495,605	869,665	1,098,381	45,7%2	420	839,87	1,506
M. gallopavo	1,043,391,504	796,560	1,000,226	45,72%	437	1,043,16	2,382
N. meleagris	1,056,996,037	798,876	1,010,756	45,83%	422	1,045,75	2,507
P. cristatus	1,355,579,690	761,396	976,691	45,62%	480	1,387,93	3,949
P. colchicus	1,123,892,910	836,354	1,074,328	44,49%	474	1,04614	2,158

After the generation of *de novo* assemblies, gene expression levels were calculated for each sample separately by mapping the reads back to the reference using RNA-Seq by Expectation-Maximization (RSEM) v 1.2.19 (Li and Dewey 2011). *De novo* RNA-Seq assemblies generate a higher number of contigs than the expected number of genes present in a genome (see **Table 3.1**). The stochastic nature of transcription results in many lowly expressed genes with possibly little biological relevance (Raj and van Oudenaarden 2008). These transcripts are unlikely to represent real coding sites and constitute noise. In order to reduce this noise and the number of contigs to a more 'realistic' number, I applied a minimum expression threshold of >2 Reads Per Kilobase of transcript per Million mapped reads (RPKM) in every tissue in at least half male/female samples, which is similar to previous studies (Harrison, et al. 2015; Wright, et al. 2015a) (box 'RPKM > 2 filter', **Figure 3.1**, **Table 3.2a**). Additionally, Ribosomal RNA (rRNA) sequences were removed as they can impact the estimation of overall expression (box 'rRNA removal', **Figure 3.1, Table 3.2a**).

RNA-Seq isoform filtering and protein prediction

After filtering genes by RPKM level, I selected the 'best' matching isoform for each Trinity 'gene' (box ''best' isoform selection', **Figure 3.1**). I defined the 'best' isoform as the isoform with the highest expression across all samples for a given species. The longest isoforms can sometimes contain chimeric sequences (Yang and Smith 2013) and by using the isoform with the highest expression, I ensured that the selected isoform is potentially biologically relevant. If two sequences showed the same expression, ties were broken using transcript length favouring the longer transcript.

Table 3.2 Filtering statistics

a) Statistics for the isoform filtering. 'Best' isoforms are those isoforms with the highest overall expression across all samples. Translated proteins are the longest those with a valid complete or partial open reading frame and a length of at least 200 amino acids (>200 AA)

***De novo* assemblies**	**RPKM 2 filtered transcripts**	**'Best' isoforms**	**Translated Proteins >200 AA**
A. cygnoides	183,677	34,873	12,455
A. platyrhynchos	181,672	30,986	13,036
M. gallopavo	202,391	38,315	11,791
N. meleagris	191,507	36,120	11,968
P. cristatus	203,468	33,079	11,127
P. colchicus	207,144	47,265	12,440

b) List of used Ensembl proteomes. For each proteome, the longest isoform was chosen and only those longer than 200 amino acids used as input data (>200 AA)

Ensembl Proteomes	**Protein coding transcripts**	**Longest isoforms**	**Longest isoforms >200 AA**
A. platyrhynchos	16,353	15,634	11,563
F. albicollis	15,983	15,303	12,367
G. gallus	16,354	15,508	12,760
M. gallopavo	16,494	14,123	11,375
T. guttata.	18,204	17,488	12,367

I used all ‘best’ isoforms as input for Transdecoder v2.0.1 (Haas, et al. 2013) to identify Open Reading Frames (ORFs) and predict protein sequences (box ‘Transdecoder translation’, **Figure 3.1**). Transdecoder.LongORFs was used with default values, except an increased minimum length of 200 amino acids (-m 200) to define an ORF followed by Transdecoder.Predict, which was used with default values to extract the most likely peptide sequences. For all species between 11,127 and 13,036 predicted expressed proteins were identified (**Table 3.2a**).

DNA-Seq data processing

In addition to the six *de novo* transcriptomes, I obtained the proteomes of all bird species available in the Ensembl v82 database (Cunningham, et al. 2015) (box ‘Ensembl proteomes’, **Figure 3.1**). Two of these are also present in the RNA-Seq dataset (turkey, *Meleagris gallopavo*; duck, *Anas platyrhynchos*), whilst one other Galliformes species (chicken, *Gallus gallus*) and two Passeriformes species (flycatcher, *Ficedula albicollis* and zebra finch; *Taeniopygia guttata*) are not. By using these additional species, comparisons between the reconstructed gene gain/loss dynamics can be made. However, because the species sets are not identical, no direct comparisons are possible, except between duck and turkey. For all proteomes, the longest isoform was used in all subsequent steps (**Table 3.2b**).

Reconstruction of gene family evolution

Orthologous Matrix (OMA) standalone v1.0.0 (Altenhoff, et al. 2015) was used for the reconstruction of the evolutionary relationships between sequences from different species (boxes ‘OMA run 1-3’, **Figure 3.1**). OMA uses all-vs-all Smith-Waterman alignments to find similar sequences, then calculates evolutionary distances before finally using a clustering step to infer the evolutionary relationships between genes (Altenhoff, et al. 2015; Roth, et al. 2008). OMA also reconstructs Hierarchical Orthologous Groups (HOGs), groups of genes that descended from a single ancestor and in a defined taxonomic range (Altenhoff, et al. 2013). HOGs were used as input for

FamilyAnalyzer (Altenhoff, et al. 2015), which uses this information to reconstruct the gene family history (stored in a HOG) consisting of loss, duplication and novel emergence events for a given taxonomic range (boxes 'Gene Family reconstruction with FamilyAnalyzer', **Figure 3.1**).

In order to identify pairs of paralogs, it is necessary to analyse the ancestral HOG composition (or gene family structure) at internal nodes within a phylogenetic tree (green box, **Figure 3.1**). HOGs are labelled using the Levels of Orthology on Trees (LOFT) (van der Heijden, et al. 2007). A HOG spanning the complete phylogenetic range is divided into labelled subgroups every time an evolutionary event (speciation or duplication) is reconstructed. FamilyAnalyzer provided these labels but did not allow direct access to the members of these subgroups at a given phylogenetic level. In order to enable access to the ancestral family size structure, I developed a patch for the tool to obtain this information, which has been integrated into the FamilyAnalyzer code base (https://github.com/DessimozLab/familyanalyzer/pull/1). The patched version of FamilyAnalyzer was used to find HOGs marked as duplicated on the terminal branch leading to a focal species and access the members of the ancestral group.

Lineage-specific paralog sequence divergence

In order to assess the divergence of potential paralog pairs, I used both protein and codon alignments (box 'Paralog sequence divergence', **Figure 3.1**). First, I performed global alignments of paralog pairs using a Needleman-Wunsch algorithm (Needleman and Wunsch 1970) for optimal, global alignments implemented in the EMBOSS v6.6.0.0 needle tool (Rice, et al. 2000) with the default EBLOSUM62 scoring matrix. I used global alignments because I expected paralog sequences to be fairly similar and because OMA considers entire proteins as evolutionary units. After aligning the paralog pairs, I extracted the percent identity and number of gaps from the needle output. Secondly, I used coding DNA sequences from each family-structure filtered HOG to create codon alignments with Muscle v3.8.31 (Edgar 2004) and TranslatorX v1.1

(Abascal, et al. 2010). For all codon alignments, a distance matrix was calculated using Biopython v1.66 (Cock, et al. 2009). This distance matrix was subsequently used to create gene trees using the neighbour joining method (Saitou and Nei 1987). Trees were inspected manually to find those gene trees in which paralogs form a monophyletic group.

Results

Are inferred evolutionary events from RNA-Seq and DNA-Seq comparable?

I first assessed the accuracy of paralog prediction from RNA-Seq data by comparing inferred proteins from the RNA-Seq based *de novo* assemblies and DNA-Seq based estimates of proteomes. If the transcriptome assembly quality is comparable to the DNA-Seq based assemblies, it is expected that general patterns of gene loss, duplication, and gain will be similar. OMA runs were repeated twice (see **Figure 3.1**), once using only the *de novo* RNA-Seq assemblies (**Figure 3.2a**) and once using only the DNA data for the Ensembl species (**Figure 3.2b**). The OMA HOGs were used to reconstruct gene losses, gains and duplications (see Methods). The species selection in both runs is not identical and direct comparisons are only possible in duck and turkey (**Figure 3.2a,b**). However, comparisons of rates across the phylogeny provide some insight into the performance of the method.

First, I compared the distributions of events across the phylogenies using absolute numbers. RNA-Seq based assemblies contain significantly more gains compared to the DNA-Seq assemblies ($P < 0.05$, $T = -2.63$, Welch's T-Test; **Figure 3.3a**). In addition to a large number of gains in RNA-Seq based assemblies, the number of losses is also significantly higher ($P < 0.05$, $T = -2.95$, Welch's T-Test; **Figure 3.3a**). The inferred duplication rate is similar between RNA- and DNA-Seq based assemblies ($P = 0.85$, $T = -0.80$, Welch's T-Test; **Figure 3.3a**). However, branch lengths are significantly shorter in the RNA-Seq based tree ($P < 0.05$, $T = -2.83$, Welch's T-Test; **Figure 3.4**), which could bias analyses when using absolute numbers. I repeated the analysis using gain, loss and duplication rates (absolute number/branch length). The reconstructed gene gain rate (duplications + singletons) for DNA data is significantly lower compared to RNA-Seq data ($P < 0.05$, $T = -3.38$, Welch's T-Test; **Figure 3.3b**), as is the loss rate ($P < 0.05$, $T = -5.46$, Welch's T-Test; **Figure 3.3b**). The difference in duplication rate is marginally non-significant but shows the same trend ($P = 0.072$, $T = -1.93$, Welch's T-Test; **Figure 3.3b**). These results suggest that the gene gain and loss dynamics appear to be different between RNA-Seq and DNA-Seq data but that inferences of duplication events are potentially comparable between both approaches.

Can RNA- and DNA-Seq data be combined to improve paralog detection?

I combined RNA-Seq data and DNA-Seq in a third OMA run to explore how these data interact and whether the combination of both data types could increase the power of the gene family reconstruction. For duck and turkey, protein sequences derived from DNA-Seq and RNA-Seq were provided as separate input files. As these sequences stem from the same species, they are expected to form sister taxa in the phylogeny with an internal ‘pseudo’ node. The OMA estimated phylogeny confirms this prediction (cladogram shown in **Figure 3.5**) and provides a topology that is identical to a taxon-restricted phylogeny provided by the BirdTree project (Jetz, et al. 2012). Including both data types for duck and turkey permits an assessment of the completeness of transcriptome analyses, the proportion of mistranslated proteins and falsely-identified duplication events.

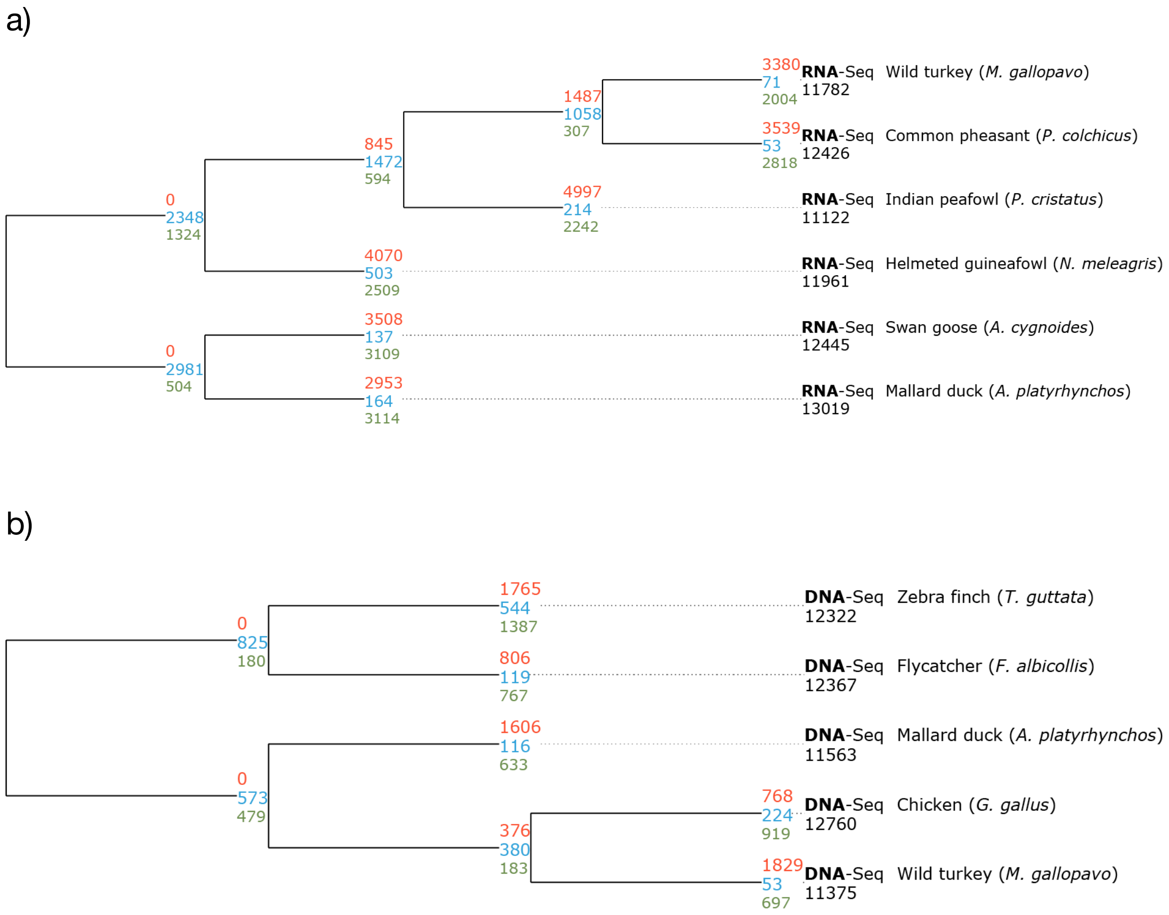

a)
3380
71
2004
RNA-Seq Wild turkey (M. gallopavo)
11782
1487
1058
307
3539
53
2818
RNA-Seq Common pheasant (P. colchicus)
12426
845
1472
594
4997
214
2242
RNA-Seq Indian peafowl (P. cristatus)
11122
0
2348
1324
4070
503
2509
RNA-Seq Helmeted guineafowl (N. meleagris)
11961
3508
137
3109
RNA-Seq Swan goose (A. cygnoides)
12445
0
2981
504
2953
164
3114
RNA-Seq Mallard duck (A. platyrhynchos)
13019
b)
1765
544
1387
DNA-Seq Zebra finch (T. guttata)
12322
0
825
180
806
119
767
DNA-Seq Flycatcher (F. albicollis)
12367
1606
116
633
DNA-Seq Mallard duck (A. platyrhynchos)
11563
0
573
479
768
224
919
DNA-Seq Chicken (G. gallus)
12760
376
380
183
1829
53
697
DNA-Seq Wild turkey (M. gallopavo)
11375

Figure 3.2 Cladogram of HOG family reconstruction for separate RNA-Seq (a) and DNA-Seq (b) assemblies. Losses (red), duplications (blue) and gains (green) are shown on internal nodes and leaves, but always refer to events that occurred on the branch leading up to this node. Numbers represent the combined count of evolutionary events across all HOGs present on that branch. Black numbers on leaves show the number of used protein sequences.

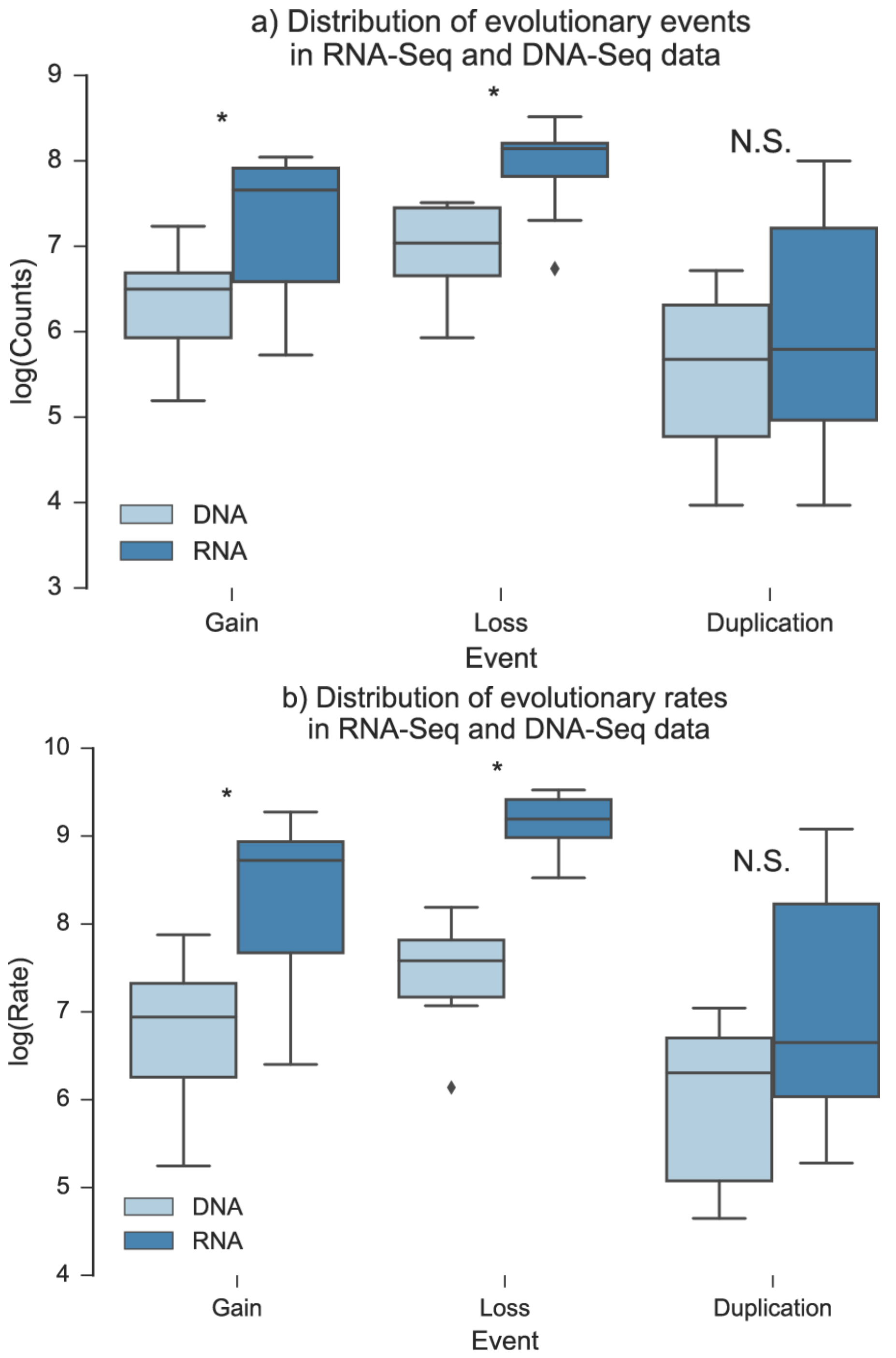


Figure 3.3 Comparison of the distribution of predicted gain, loss and duplication events across all branches in the RNA-Seq (dark blue, information from **Figure 3.2a**) and DNA-Seq (light blue, information from **Figure 3.2b**) OMA runs. a) Absolute natural *log* transformed counts were used for all events b) Natural *log* transformed gain, loss and duplication rates (number of events/branch length). Significance levels are indicated as stars (* $P < 0.05$, ** $P < 0.001$, *** $P < 0.0001$), differences between distributions were tested using Welch's T-Test.

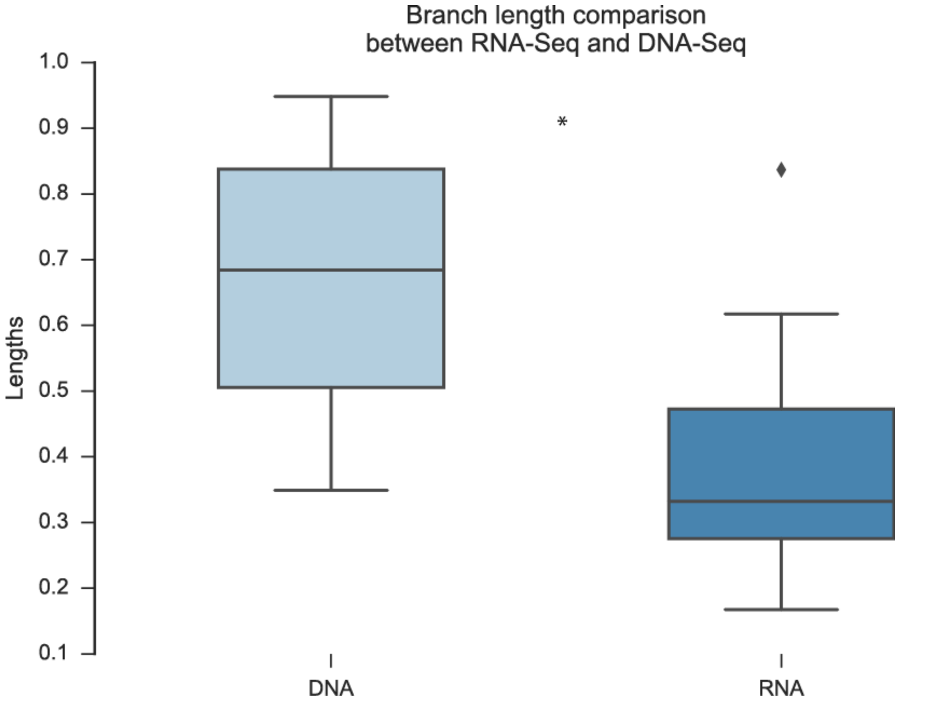


Figure 3.4 Distribution of branch lengths (expected number of substitutions per site) in the RNA-Seq (dark blue) and DNA-Seq (light blue) OMA runs. Significance levels are indicated as stars (* *P* < 0.05*,* ** *P* < 0.001, *** *P* < 0.0001), differences between distributions were tested using Welch's T-Test.

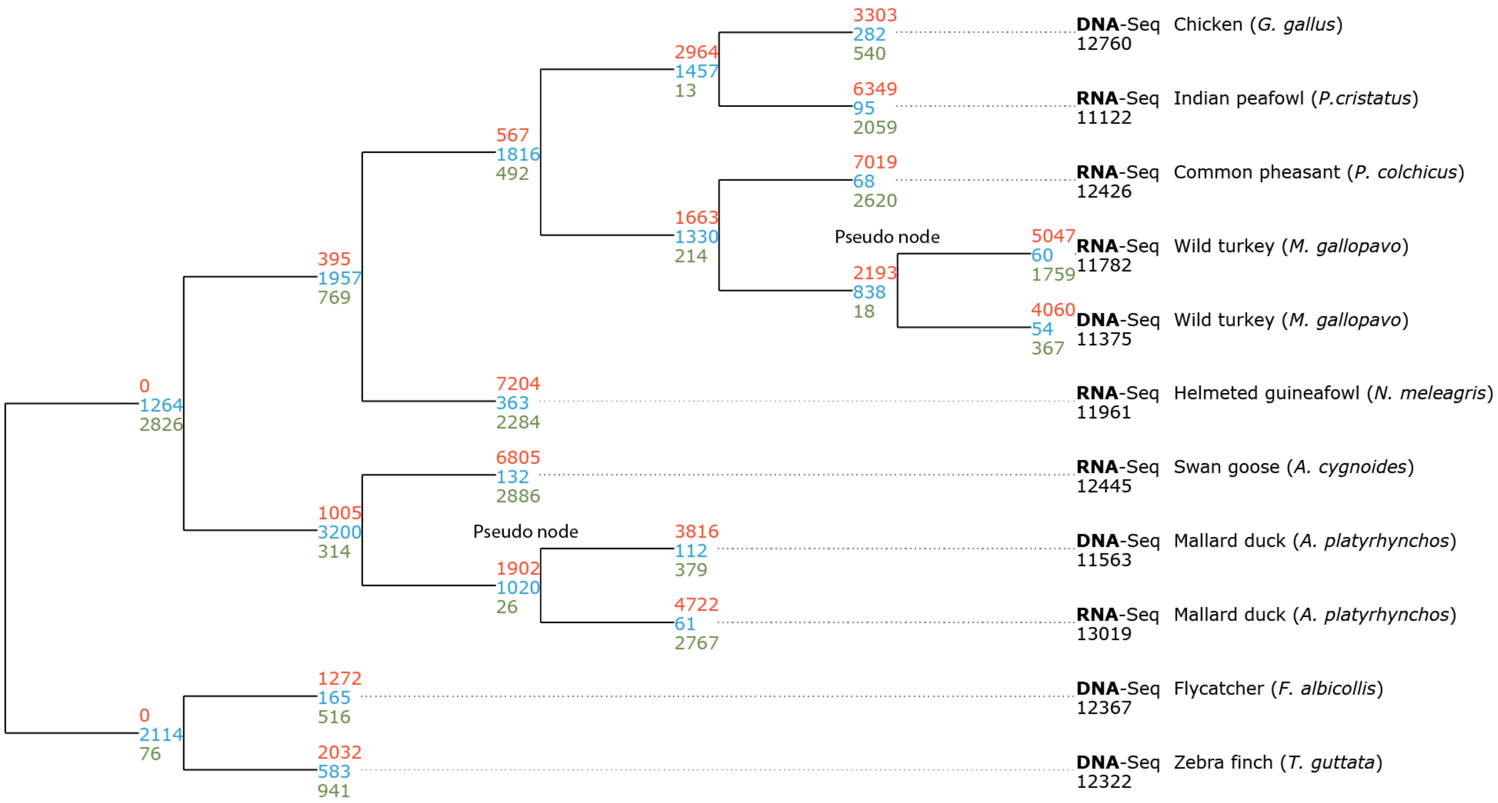

3303
282
540
DNA-Seq Chicken (G. gallus)
12760
2964
1457
13
6349
95
2059
RNA-Seq Indian peafowl (P.cristatus)
11122
567
1816
492
7019
68
2620
RNA-Seq Common pheasant (P. colchicus)
12426
1663
1330
214
Pseudo node
5047
60
1759
RNA-Seq Wild turkey (M. gallopavo)
11782
395
1957
769
2193
838
18
4060
54
367
DNA-Seq Wild turkey (M. gallopavo)
11375
0
1264
2826
7204
363
2284
RNA-Seq Helmeted guineafowl (N. meleagris)
11961
6805
132
2886
RNA-Seq Swan goose (A. cygnoides)
12445
1005
3200
314
Pseudo node
3816
112
379
DNA-Seq Mallard duck (A. platyrhynchos)
11563
1902
1020
26
4722
61
2767
RNA-Seq Mallard duck (A. platyrhynchos)
13019
1272
165
516
DNA-Seq Flycatcher (F. albicollis)
12367
0
2114
76
2032
583
941
DNA-Seq Zebra finch (T. guttata)
12322

Figure 3.5 Cladogram (branch lengths do not reflect evolutionary distance) of all used RNA-Seq and DNA-Seq assemblies. Losses (red), duplications (blue) and gains (green) are shown on internal nodes and terminal branches, but always refer to events that occurred on the branch leading up to this node. Gains on terminal branches represent singletons. Numbers represent the combined count of evolutionary events across all HOGs present on that branch. Black numbers on leaves show the number of used protein sequences.

The 16,546 pairwise ortholog groups calculated by OMA were used to infer 22,126 Hierarchical Orthologous Groups HOGs (see Methods). Across the phylogeny, significantly higher numbers of duplications are reconstructed on internal branches, with lower numbers on terminal branches ($P < 0.001$, $T = 5.83$, Welch's *T*-Test; **Figure 3.5**). Conversely, losses ($P < 0.05$, $T = -3.32$, Welch's *T*-Test; **Figure 3.5**) and gains ($P < 0.05$, $T = -2.61$, Welch's *T*-Test; **Figure 3.5**) are significantly more common on terminal branches compared to internal branches. However, predicted loss, gain and duplication events are common across all branches (**Figure 3.5**), which may be an indication of highly fragmented assemblies (Altenhoff, et al. 2015).

I used the 'pseudo' nodes to directly assess the effects of combining different data types. If the transcriptome and genome both contain identical information, no gain, loss or duplication events are expected on 'pseudo' branches. However, the number of inferred proteins between both datatypes is not identical and a low amount of reconstructed events may be expected. First, I compared the proportion of gains between the RNA-Seq and DNA-Seq assemblies and found that both in turkey ($P< 0.0001$, Odds ratio=5.26; Fisher's Exact Test) and in duck ($P< 0.0001$, Odds ratio=7.97; Fisher's Exact Test) the proportion of gains is significantly higher in RNA-Seq assemblies compared to DNA-Seq assemblies. In duck, ca. 21% (2767/13019) and in turkey ca. 15% (1759/11782) of all genes are reconstructed as gained on the terminal branch. This is indicative of a large amount of noise in the RNA-Seq assemblies and in line with the high numbers of losses reconstructed on terminal branches. The proportion of duplication events is similar in both turkey assemblies ($P > 0.05$, Odds ratio=1.07; Fisher's Exact Test) but significantly higher in the duck RNA-Seq assembly compared to the DNA-Seq assembly ($P < 0.0001$, Odds ratio=0.48.; Fisher's Exact Test). This could indicate a false identification of isoforms as paralogs in the RNA-Seq assembly. Taken together, these results suggest that a combined RNA-Seq and DNA-Seq contains large amounts of noise and that additional filtering is required.

Can lineage-specific paralogs be filtered using ancestral family structure?

The high number of gains and losses predicted on all branches (**Figure 3.5**) make it necessary to filter the data further in order to reliably detect paralogs. I used a filter that depends on the availability of the ancestral gene family structure composition, a feature that I patched into the FamilyAnalyzer toolkit (see Methods). The ancestral gene family structure is used to distinguish families that are marked as duplicated and are affected by differential gene loss from those families that are not. In order to obtain a reliable set of lineage-specific paralogs, I inspected the ancestral family composition for all HOGs marked as duplicated on terminal branches. Only those families where the ancestral family structure contained two copies in the focal species and only one gene from the most closely related outgroup species were kept. For example, a HOG marked as duplicated on the branch leading to the peafowl should contain two paralogs in the peafowl and only one ortholog in chicken. Using this criterion, I extracted candidate HOGs for all six species with available RNA-Seq data. For turkey and duck, I required duplicated HOGs to contain two paralogs in the RNA-Seq and DNA-Seq assemblies. Using this filter, nine lineage-specific duplications in turkey, five in duck, 15 in the peafowl, 15 in the pheasant, 28 in the swan goose and 18 in the helmeted guinea fowl were retained (**Table 3.3**). The number of paralog pairs in turkey and duck is likely lower because the filtering required two sequences to be present in the DNA-Seq and RNA-Seq assemblies. In comparison to unfiltered estimates, this set of paralogs is much smaller; however, further verification is needed to ensure that these paralogs are not false-positives and that the family-structure filtering was successful.

Table 3.3 Retained lineage-specific paralog pairs after each filtering step. Numbers in brackets are the estimated duplication rate per million years using median divergence times obtained from TimeTree (Hedges, et al. 2015)

RNA-Seq Species	Unfiltered	Family structure filtered	> 10%and<100% protein seq. identity	Paralog pairs form a monophyl. group
A. cygnoides	132 (4.44)	28 (0.94)	23 (0.77)	2 (0.07)
A. platyrhynchos	61 (2.05)	5 (0.17)	5 (0.17)	0 (0.00)
M. gallopavo	54 (1.68)	9 (0.28)	9 (0.28)	3 (0.09)
N. meleagris	363 (7.53)	18 (0.37)	11 (0.23)	2 (0.04)
P. cristatus	95 (2.71)	15 (0.43)	15 (0.43)	1 (0.03)
P. colchicus	68 (2.11)	15 (0.47)	10 (0.47)	2 (0.06)

In order to ensure that the family filtered paralog pairs are not false-positives, I created global protein sequence alignments to assess sequence identity between paralog pairs. I expected that recent, lineage-specific paralogs would have a relatively high protein sequence identity. However, the percentage identity of alignments between paralogs ranges from less than 5% to exactly 100% (mean identity = 50.47%, standard deviation = 29.49; **Figure 3.6**). The similarity distribution is skewed towards lower identity and 7.8% (7/90) of paralog pairs have a sequence identity of less than 10%. For paralog pairs with sequence identity lower than 10%, manual inspection of the alignments revealed that in all cases only a short region of both sequences aligns with high identity, and this region is surrounded by complete mismatches. At the other end of the distribution, 10% (9/90) of all paralog pairs have 100% protein sequence identity. Manual inspection revealed that all cases originated from different ‘genes’ and ‘best’ isoforms in the same Trinity inferred transcript clusters. It is possible that these cases constitute very recent duplications, but it is more likely that these paralogs are artefacts produced by Trinity, potentially due to difficulties with isoform resolution.

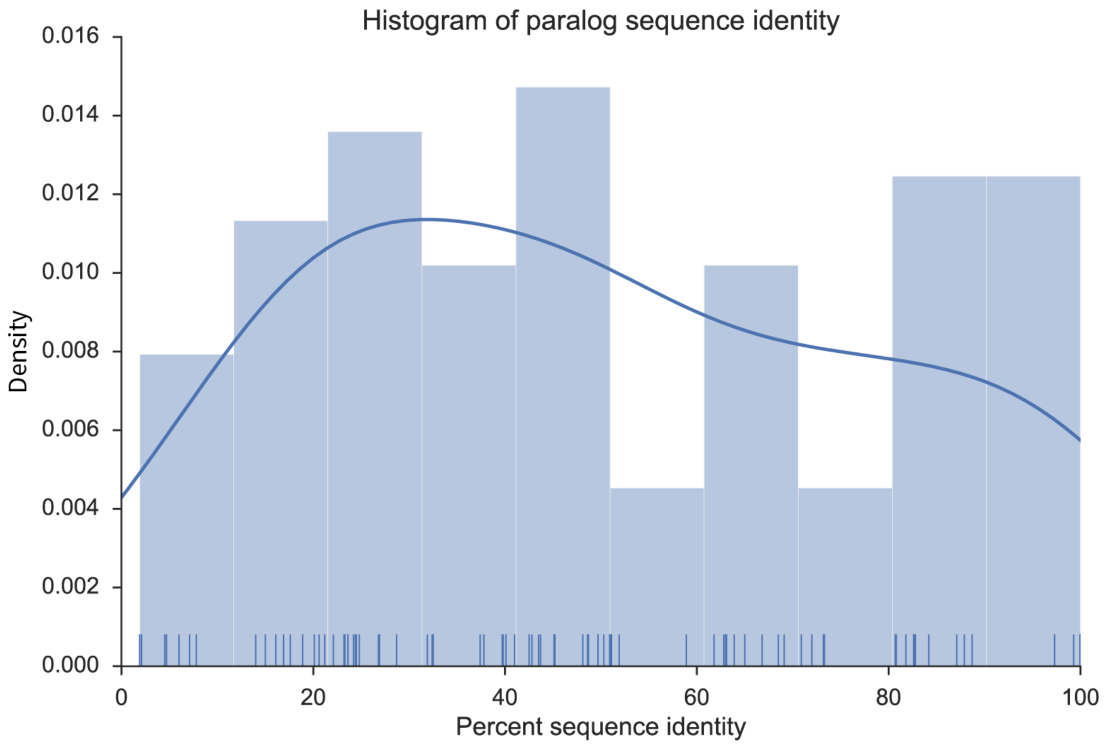


Figure 3.6 Histogram of the percent protein sequence identity between lineage-specific pairs of paralogs across all six RNA-Seq based bird species. Sequence identity was obtained from global Needleman-Wunsch alignments. Rugs indicate the exact percent identity of each data point; rugs can overlap.

The low protein sequence identity of some paralogs could be due to adaptive divergence. However, recently duplicated lineage-specific paralogs are expected to form monophyletic groups in comparison to the non-duplicated ortholog in a closely related taxon, as it is unlikely that these sequences had enough time to diverge sufficiently. In order to test this hypothesis, I constructed gene trees using codon alignments for all pairs of paralogs in every species (see Methods). I used nucleotide sequences for the reconstruction of gene trees because they contain many synonymous sites, which are expected to be dominated by drift and are more reliable for inferring evolutionary relationships. With the exception of three pairs in turkey, and two pairs in both the swan goose and Indian peafowl, paralogs did not show the expected pattern of relatedness. No paralogs remained to be assessed for duck (**Table 3.3**). This suggests that either these paralogs diverged extremely quickly, or that the sequence pairs are falsely identified as paralogs. All of the pairs that formed monophyletic groups had a protein sequence identity of at least 31% and eight of the pairs align with more than 50% sequence identity. Given the low protein sequence identity between many other paralogs, the large amount of noise in the RNA-Seq data and the young age of these lineage-specific paralogs, it is likely that the majority of inferred paralogs are false positives, even after strict filtering.

Is the duplication rate of inferred lineage-specific paralogs comparable to genomic estimates?

In order to assess whether the inferred paralog pairs are true paralogs, I compared the inferred duplication rates to previous estimates based on genomic data. Avian genomes are thought to be more stable compared to mammals (Hillier, et al. 2004; Jarvis, et al. 2014), with a relatively low duplication rate of ca. 0.28 single gene duplications per million years (Toups, et al. 2011), based on combined estimates of retrotransposition (21 genes) and DNA-mediated events (8 genes). Before filtering, the duplication rates on terminal branches are extremely high, ranging from 1.68 duplications per million years in turkey to 7.53 in the helmeted guineafowl (**Table 3.3**), on average ca. 12 times higher than estimated using genome data (Toups, et al. 2011). Even after filtering by family structure, the duplication rate in lineages with only RNA-Seq data available are up to three times higher in comparison to estimates by Toups, et al. (2011) (**Table 3.3**). In duck the duplication rate from the internal node to the ‘pseudo-node’ is lower; in turkey it is similar to genomic estimates. This suggests that RNA-Seq based assemblies overestimate the number of lineage-specific duplications and in combination with the large differences in duplication rate between species, limit the credibility of the reconstruction results.

Can strictly filtered paralogs be confirmed using Ensembl data?

The low rates of protein identity, unexpected gene tree topologies, and higher-than-expected duplication rates suggest that even strictly filtered sets of paralogs contain a large number of false-positives. For duck and turkey, data from RNA-Seq and DNA-Seq assemblies were used in conjunction, allowing the validation of gene duplications with Ensembl Compara (Cunningham, et al. 2015), which uses only genomic data. Family-structure filtered HOGs contain two RNA-Seq based sequences and two DNA-Seq based sequences obtained from Ensembl, plus one outgroup sequence. Every pair of Ensembl sequences could be confirmed in the Ensembl Compara database; however, none of the Ensembl paralogs were listed as lineage-specific. Instead, all duplication events were much older, and preceded the origin of the avian clade (**Table 3.4**). The misplacement of duplication events is potentially the result of missing data in the RNA datasets, which may vary stochastically across species, and pushes duplication events closer to the tips of the tree as the probability of lineage specific gains outweighs that of multiple losses. This indicates that OMA is able to infer paralogs correctly when DNA-Seq data is used but that the branch assignment is likely incorrect.

Table 3.4 Pairs of Ensembl paralogs in family-structure filtered HOGs in duck and turkey. Data obtained from Ensembl Compara (Cunningham, et al. 2015).

Species	Paralog 1	Paralog 2	Ancestral taxonomy
A. platyrhynchos	ENSAPLG00000010742	ENSAPLG00000006733	Euteleostomi
	ENSAPLG00000007753	ENSAPLG00000004048	Euteleostomi
	ENSAPLG00000009198	ENSAPLG00000002664	Vertebrata
	ENSAPLG00000001211	ENSAPLG00000003661	Euteleostomi
	ENSAPLG00000003580	ENSAPLG00000009099	Vertebrata
M. gallopavo	ENSMGAG00000012781	ENSMGAG00000004944	Vertebrata
	ENSMGAG00000014650	ENSMGAG00000010624	Vertebrata
	ENSMGAG00000002202	ENSMGAG00000008378	Euteleostomi
	ENSMGAG00000007562	ENSMGAG00000010129	Euteleostomi
	ENSMGAG00000014131	ENSMGAG00000003655	Euteleostomi
	ENSMGAG00000004136	ENSMGAG00000015066	Euteleostomi
	ENSMGAG00000009588	ENSMGAG00000000915	Euteleostomi
	ENSMGAG00000006850	ENSMGAG00000015234	Chordata
	ENSMGAG00000006598	ENSMGAG00000008568	Euteleostomi

Discussion

De novo RNA-Seq assemblies have facilitated a burst in molecular analyses in ‘non-model’ organisms (Ekblom and Galindo 2011; Wang, et al. 2009), where genomic resources are unavailable. However, many comparative RNA-Seq studies are restricted to comparisons of orthologs, which limits the analyses to a set of genes that is more likely to maintain a similar overall function in different species (Koonin 2005). For this reason, new methods are needed that are able to accurately reconstruct the evolutionary history of gene families inferred solely based on RNA-Seq data to enable comparative analyses of paralog sequence and gene expression divergence. Given the important role gene duplication plays in adaptation and the evolution of phenotypic novelty (Ohno 1970; Zhang 2003), developing such an approach is crucial to fully understanding how selection shapes the genome, and through it, organismal diversity.

In this chapter, I explored the potential use of *de novo* transcriptome assemblies using RNA-Seq for inferring orthologs and paralogs, with the aim of using this information to investigate the divergence of gene expression and gene sequence between paralogs. I used OMA with protein sequences predicted from RNA-Seq to reconstruct gene family evolution. OMA is well-suited to this task because it supports the reconstruction of HOGs solely based on the orthology graph, which eliminates the need for additional tools to perform gene and species tree reconciliation (Altenhoff, et al. 2013; Altenhoff, et al. 2015). The only input required to run this analysis is a set of protein sequences from different species, independent of their original source (RNA- or DNA-Seq). The resulting HOGs can be used to detect gene duplications across the phylogeny and, in combination with FamilyAnalyzer, allow for the convenient data exploration and analysis. My results, however, indicate several methodological issues that limit the applicability of the developed bioinformatics pipeline for the detection of paralogs using *de novo* RNA-Seq data.

Family structure filtering and duplication rate comparisons

I patched the FamilyAnalyzer toolkit to enable direct access to the ancestral gene family composition. This information was then used to filter HOGs based on an expected gene family structure (see Methods, **Figure 3.1**). The ancestral family size composition revealed that differential gene loss has likely affected the reconstruction of duplication events because in many families marked as duplicated, only one sequence from each species is present. If a single gene duplication occurred in the common ancestor of two species, it generates two outparalogs. Differential gene loss after the speciation event can then lead to a situation where outparalogs are marked as orthologs. OMA and the algorithm to compute HOGs try to detect these 'pseudo' orthologs (Altenhoff, et al. 2013) and remove them. However, the high number of differential gene losses caused by noisy RNA-Seq assemblies could falsely lead to the inference of too many duplication events. This also constitutes a potential explanation for the elevated rates of loss and gain in RNA-Seq based assemblies (**Figure 3.3**).

Only a subset of families was kept and used for further analyses; however, even after strict filtering the inferred gene duplication rate (**Table 3.3**) show that the RNA-Seq based estimates are two to four times higher compared to genomic estimates (Toups, et al. 2011). Avian genomes are more stable and compact in comparison to other amniotes (Hillier, et al. 2004; Organ, et al. 2007), with less genetic gain and loss compared to mammals (Jarvis, et al. 2014). In contrast to RNA-Seq based estimates, duck and turkey duplication rates incorporate genomic data and the duplication rate is similar to previous estimates. Although some variation in gene duplication rate is possible, the consistent observation that RNA-Seq derived estimates of duplication rate are inflated is likely to be a consequence of low RNA-Seq assembly quality and not an accelerated rate of gene duplication.

Paralog pairs show low sequence identity and unexpected gene tree topology

When analysing the protein sequence alignments of family-structure paralog pairs, ca. 20% of alignments showed either extremely low sequence identity (<10%) or showed that the protein sequences of paralogs were identical. All identical sequences were part of the same Trinity read cluster, which are assembled separately, and group together predicted ‘genes’ and isoforms. When different genes from the same cluster contain exactly the same open reading frame, it is likely that these are not paralogs, but redundant Trinity artefacts. OMA performs local Smith-Waterman alignments, and manual inspection of alignments revealed that many paralog pairs only share a relatively short region with very high identity, surrounded by areas with very low identity. This is consistent with low assembly quality and indicative of erroneous or incomplete open reading frame prediction in the RNA datasets by Transdecoder.

One possible alternative explanation for this pattern is that OMA’s default length tolerance ratio (LengthTol * min(length(s1), length(s2))) (Roth, et al. 2008) of 0.61 could be insufficient for RNA-Seq data. However, benchmark analyses of OMA indicated that OMA’s performance plateaus between a length tolerance threshold of 0.61 and 0.9 (Roth, et al. 2008). This suggests that a stricter cut-off criterion would not resolve these issues, although as OMA was originally designed for DNA-Seq data, it remains possible that the length tolerance threshold performs differently on mixed RNA/DNA datasets.

Over larger phylogenetic distances, paralog pairs are expected to diverge either due to drift or divergent selective pressures (see Hahn 2009; Innan and Kondrashov 2010). Consequently, at least some of the protein sequence divergence could be due to adaptive processes. However, I expected that recently duplicated, lineage-specific paralogs would form a monophyletic group when aligned with orthologs in sister taxa. For these analyses, I used codon alignments because it is likely that synonymous variation still captures information about the evolutionary relatedness of gene sequences, provided the exon structure is largely conserved. Gene trees of lineage-specific paralog pairs and their orthologs in sister species showed that only 1-3 pairs of paralogs out of 90 followed this pattern. This is consistent with the finding that many predicted paralogs have low sequence identity.

Taken together, these results limit the credibility of the inferred paralog pairs using the bioinformatics pipeline I developed. After family-structure filtering and the removal of paralogs with unexpected gene tree topologies, only one or two paralog pairs per species were retained. Given the small number of retained genes, it is likely that a change of parameters could result in different sets of genes, with little overlap between runs. Unfortunately, these analyses therefore did not provide the basis for a comparative analysis of general patterns of gene sequence and expression divergence following duplication events.

An explanation: variation in gene model prediction

In this analysis, *ab initio* protein translation was performed using Transdecoder (Haas, et al. 2013), which makes predictions solely based on the transcript sequence and does not rely on any sequence alignments. The identification of open reading frames and subsequently protein-coding sequences is challenging and often results in suboptimal annotation quality (Yandell and Ence 2012). Terminal branches contain a large number of novel singletons, proteins that could not be aligned to any other sequence. This could be caused by interspecific differences in the temporal and spatial dynamics of transcription; however, given the comparable sampling across species in combination with the high numbers of gains in all *de novo* assemblies, this is likely to have only a minor effect. Alternatively, the *ab initio* predicted proteins from *de novo* assemblies could be incomplete and contain a high number of falsely translated proteins. This is a central issue because differences in gene model quality between species make it extremely difficult to accurately infer relationships between genes. For example, incorrect gene model predictions can lead to exons being missed or the retention of non-coding DNA in the predicted protein coding region (Drăgan, et al. 2016). The pipeline developed here does not account for these differences and weighs each assembly equally. This introduces a bias to the reconstruction of HOGs and reduces confidence in the reconstruction of lineage-specific paralogs. The recent emergence of methods that quantify confidence in gene models, such as GeneValidator (Drăgan, et al. 2016) could be used to identify problems with protein-coding gene predictions before running OMA to identify HOGs.

The reconstruction of gene families in general, and specifically the detection of paralogs from *de novo* transcriptome data, remains challenging. Despite a series of careful steps to filter the data, the high number of likely false-positives that remain in the dataset suggest confidence in the reconstruction is low. The higher number of likely false-positives also prevents further analyses of paralog expression divergence using gene expression estimates. Highlighting these analytical problems is nonetheless important because they demonstrate current issues and limitations of using *de novo* RNA-Seq assemblies for the reconstruction of gene family evolution. My analyses also indicate some fundamental limitations of using RNA-Seq data for comparative analyses. The spatial and temporal variance in gene expression makes it challenging to obtain comparable RNA-Seq libraries across a wide phylogenetic range. Developmental time points vary between species, and gene expression changes can occur relatively quickly. Consequently, the overlap between expressed genes may be low (Harrison, et al. 2015), which poses limitations on the usefulness of RNA-Seq data for the detection of orthologs and paralogs. For now, it may be advisable to create a set of reference genomes across species to generate comparable gene models, before conducting RNA-Seq experiments to investigate the gene expression divergence of paralogs. In the future, the availability of high-throughput and long read sequencing technologies could mitigate many of the assembly issues.

Chapter 4

The potential role of sexual selection and sexual conflict on the genomic distribution of mito-nuclear genes

The analyses presented in this chapter have been published in *Genome Biology and Evolution*:

Dean R, **Zimmer F**, Mank JE 2014. The potential role of sexual conflict and sexual selection in shaping the genomic distribution of mito-nuclear genes. *Genome Biology and Evolution* 6(5): 1096-1104, **doi**: 10.1093/gbe/evu063

Author contributions:

I designed the data analyses with Dr Rebecca Dean and wrote the paper in collaboration with Dr Rebecca Dean and Professor Judith Mank.

I contributed the bioinformatics analyses, including all BLAST-based analyses and the detection and localization of genes interacting with mitochondria.

Dr Rebecca Dean performed the statistical analyses to test for significant over- or underrepresentation of mitochondrial interacting genes on the X or Z chromosomes.

I thank Beatriz Vicoso and Doris Bachtrog for data on Schistosoma X chromosome contigs and Michael Stocks for computational advice.

Summary

Mitochondrial interactions with the nuclear genome represent one of life's most important coevolved mutualisms. In many organisms, mitochondria are maternally inherited, and in these cases co-transmission between the mitochondrial and nuclear genes differs across different parts of the nuclear genome, with genes on the X chromosome having 2/3 probability of co-transmission, compared to 1/2 for genes on autosomes. These asymmetrical inheritance patterns of mitochondria and different parts of the nuclear genome have the potential to put certain gene combinations into inter-genomic coadaptation or conflict. Previous work in mammals found strong evidence that the X chromosome has a dearth of genes that interact with the mitochondria (mito-nuclear genes, mt-N genes), suggesting that inter-genomic conflict might drive genes off the X onto the autosomes for their male-beneficial effects. Here, we developed this idea to test co-adaptation and conflict between mito-nuclear gene combinations across phylogenetically independent sex chromosomes on a far broader scale. We found that, in addition to therian mammals, only *Caenorhabditis elegans* showed an underrepresentation of mito-nuclear genes on the sex chromosomes. The remaining species studied showed no overall bias in their distribution of mito-nuclear genes. We discuss possible factors other than inter-genomic conflict that might drive the genomic distribution of mito-nuclear genes.

Introduction

The eukaryotic cell contains two distinct genomes - the nuclear and the mitochondrial – whose coordinated interactions over billions of years now represent one of life's most important coevolved mutualisms (Gillham 1994). Many gene products are encoded in the nucleus and exported to the mitochondria where they interact with other, mitochondrially-encoded, genes. Organismal fitness depends upon compatibility between nuclear and mitochondrial gene products (Meiklejohn, et al. 2013), and these interactions (hereafter 'mito-nuclear') are fundamental to eukaryotic existence and underlie key life history traits, including somatic maintenance, reproductive performance and ageing (Dowling, et al. 2008; Rand, et al. 2004).

However, because mitochondria are often maternally inherited, selection acting on these mito-nuclear interactions is asymmetrical in males and females. Mutations detrimental to males are not selected against unless they are also detrimental to females, except in exceptional cases involving non-random mating, sperm limitation, or paternal mitochondrial transmission (e.g., Hedrick 2012; Rand, et al. 2001; Unckless and Herren 2009; Wade and Brandvain 2009; Zhang, et al. 2012). In extreme cases, mitochondrial mutations that harm males can even be selected for if they benefit females. This results in a male mutational load where mutations detrimental to males are not purged from populations and accumulate across generations (Frank and Hurst 1996; Gemmell, et al. 2004). This male mutational load can be detected in the form of male-biased gene miss-expression (Innocenti, et al. 2011), reduction in male lifespan (Camus, et al. 2012) and in male fertility (Smith, et al. 2010; Yee, et al. 2013) in individuals that contain mitochondria from different populations.

Maternal inheritance of mitochondria puts mitochondrial genes in contrasting evolutionary dynamics with different parts of the nuclear genome: whereas Y chromosomes have strict paternal transmission, autosomes are equally transmitted through males and females, and X chromosomes spend twice their time in females compared to males. This sexual asymmetry across the genome might set the scene for inter-genomic coadaptation or conflict. On the one hand, it is expected that beneficial gene combinations are facilitated if genes that interact with the mitochondria are on the X chromosome. The X chromosomes in mammals and *Drosophila* have been shown to be feminized for gene expression (Khil, et al. 2004; Meisel, et al. 2012), and X-linked genes are co-transmitted with mitochondrial genes through the female 2/3 of the time. Under such a scenario - with inter-genomic coadaptation driving the distribution of genes that interact with mitochondria – an over-representation of mito-nuclear genes on the X may be expected (Rand, et al. 2001; Wade and Brandvain 2009; Wade and Goodnight 2006). On the other hand, the accumulation of mutations that are detrimental to males, referred to as male-biased mitochondrial mutational load, might be ameliorated if genes that interact with the mitochondria move off the X, onto parts of the genome with equal (or even male-biased) transmission. If conflict drives the distribution of mito-nuclear genes, an underrepresentation of genes that interact with the mitochondria on the X chromosome would be expected (Drown, et al. 2012; Rice 1984; Werren 2011).

Converse patterns are expected for Z chromosomes in female-heterogametic species. ZW systems often show reverse patterns for sexual conflict scenarios since the Z is masculinized (Wright, et al. 2012) while the X is feminized for gene expression. This potentially results in an underrepresentation of mito-nuclear genes on the Z chromosome since mitochondria are co-transmitted with Z chromosomes only 1/3 of the time. Alternatively, because the Z and mitochondria can never be transmitted through males, it is possible that there is no expected bias on Z chromosomes with regard to mito-nuclear genes (Drown, et al. 2012). Finally, it has also been suggested that the Z chromosome might be enriched for mito-nuclear genes due to sexual selection in males (Hill and Johnson 2013).

These predictions for the distribution of mito-nuclear genes are predominantly based on probabilities of co-inheritance of mitochondria with different parts of the nuclear genome and do not take into account more complex processes such as linkage patterns of genes interacting with mitochondria. Empirical evidence for mito-sex chromosome interactions is not consistent. Some experimental evidence suggests genes on the X chromosome interact with mitochondrial genomes in *Drosophila* (Rand, et al. 2001), whereas other assessments failed to detect mito-autosomal interactions (Clark 1985; Clark and Lyckegaard 1988). Consistent with the predictions of inter-genomic conflict, a strong underrepresentation of mitochondrial genes on the X chromosome was found across a range of mammal species (Drown, et al. 2012). However, the dataset used by Drown, et al. (2012) is phylogenetically non-independent, as the X chromosomes are orthologous and their gene contents are highly conserved across the therian mammals (Veyrunes, et al. 2008), therefore the universality of the dearth of mitochondrial genes on the X remains largely unexplored.

Here, we test the universality of predictions of mito-nuclear coadaptation and conflict by exploring the genomic distribution of genes that interact with the mitochondrial genome. We extend previous studies by exploring these interactions on a broad scale, incorporating multiple examples of male- and female-heterogamety in species with independent origins of their sex chromosomes.

Material and Methods

Detection and localization of genes interacting with mitochondria

In order to expand the analysis to species with less complete genome annotations, I modified the protocol from Drown, et al. (2012) to compare the chromosomal distribution of genes that interact with the mitochondria across a range of species with phylogenetically independent sex chromosomes. In the first step, the proteomes for the several therian mammals (*Bos taurus, Callithrix jacchus, Canis familiaris, Gorilla gorilla, Homo sapiens, Macaca mulatta, Equus caballus, Oryctolagus cuniculus, Pongo abeloo, Rattus norvegicus, Sus scrofa and Monodelphis domestica*), the monotreme *Ornithorhynchus anatinus*, three birds (*Gallus gallus*, *Meleagris gallopavo* and *Taeniopygia guttata*), the fish *Gasterosteus aculeatus*, *Drosophila melanogaster* and *Caenorhabditis elegans* from Ensembl v71(Flicek, et al. 2014) were obtained. In order to increase the number of independently-evolved sex chromosomes, the proteomes for *Tribolium castaneum*, *Bombyx mori* and *Schistosoma mansoni* from Ensembl Metazoa v18 (Kersey, et al. 2012) were also obtained.

Because genome and gene ontology annotation quality varies across species, I used a reciprocal best BLAST (Altschul, et al. 1990) hit (rBBH) approach to find one-to-one orthologs between the well-annotated *Mus musculus* mito-nuclear genes and the other species using the catalogue of genes with mitochondrial annotation (mito-nuclear genes) in the Gene Ontology (Ashburner, et al. 2000) ID 0005739 for *Mus musculus* using BIOMART (Durinck, et al. 2005) from Ensembl v71 (Flicek, et al. 2014). This approach relies on the high level of conservation of mitochondrial gene function (Jafari, et al. 2013; Lotz, et al. 2013). To verify that rBBH is appropriate for mito-nuclear genes, we compared the list of genes obtained through rBBH with the list of mitochondrially annotated genes using Gene Ontology term GO:0005739 in Biomart for *D. melanogaster* and *C. elegans* - two species with more complete gene annotation. We found that out of the 522 *D. melanogaster* GO:0005739 genes, 66% (345/522) were also identified as mito-nuclear by the rBBH. Of the 251 *C. elegans* GO:0005739 genes, only 7% (18/251) were also identified through the rBBH. This suggests, that while rBBH is useful for detecting mito-nuclear orthologs (comparable with computational annotation of GO terms), this approach may miss or incorrectly classify some of the mito-nuclear genes across distantly related species.

In order to account for clade-specific differences, I conducted two further analyses. First, I repeated the rBBH analysis, using Biomart to identify mito-nuclear GO:0005739 genes for *D. melanogaster* and *C. elegans* in addition to *M. musculus.* Because these are relatively well annotated genomes, I used them as clade-specific reference species in order to reduce taxonomic distance. Therefore, I used (a) *M. musculus* mito-nuclear genes as the reference for other vertebrates (Theria, *O. anatinus, G. aculeatus*, and Aves), (b) *D. melanogaster* mito-nuclear genes as the reference set for other insects (*T. castaneum* and *B. mori*) and (c) *C. elegans* mito-nuclear genes for the entozoans (with *S. mansoni*). Secondly, I also present results using just Biomart GO term annotations for those species where gene products have been annotated.

For the rBBH analysis, I used the longest protein isoform and only considered hits when the BLASTP (Altschul, et al. 1990) E-value was below 10^{-7}. In the second rBBH analysis, also using *D. melanogaster* and *C. elegans* as reference points, I used a more stringent E-value threshold of 10^{-10}; hits were then ordered by bitscore and a rBBH was accepted only when the best hit had a sequence identity larger than 30%. After the rBBH analyses, I determined the chromosomal location for mouse mito-nuclear orthologs in each species. The *S. mansoni* locations are based on Vicoso and Bachtrog (2011), *B. mori* positions were extracted from KAIKObase version 3.2.1 (Shimomura, et al. 2009), *T. castaneum* are based on Ensembl Metazoa v18 (Kersey, et al. 2012) and all other locations are based on Ensembl v71 (Flicek, et al. 2014).

As a result, three lists of nuclear genes with mitochondrial annotation and their chromosomal locations were created: (1) using direct GO annotation (only in *M. musculus*) or based on orthology predictions (all other species), (2) based on direct GO annotation (*M. musculus*, *D. melanogaster* and *C. elegans*) or based on orthology predictions using the closest relative from these three species, and (3) based on direct GO annotation, just for *O. anatinus* and *G. aculeatus* (*S. mansoni*, *T. castaneum* and *B. mori* are not available in Ensembl, and Theria and Aves have previously been reported using this approach by Drown, et al. (2012)).

Statistical analysis

In order to avoid problems with phylogenetic non-independence, we combined all species that share the same orthologous sex chromosome into a single data point (i.e. the therian mammals were grouped together, as were the birds). We then compared the density of mito-nuclear genes on the sex chromosomes and the autosomes relative to the expected gene density based on the total number of mitochondrial annotated genes. For *D. melanogaster,* each Muller element (X, 2L, 2R, 3L, 3R, 4) was treated as a separate chromosome. The expected gene count per chromosome was calculated as the total number of mito-nuclear genes multiplied by the proportion of all annotated genes on each chromosome. The bias of mito-nuclear genes was the ratio of the observed number of mito-nuclear genes on a chromosome to the expected count, where an over-representation is a bias > 1 and an underrepresentation is a bias < 1. In *G. aculeatus*, we also included the neo-sex chromosome (Kitano, et al. 2009; Natri, et al. 2013), as well as the *D. melanogaster* ancient-sex chromosome, which displays many properties of an X chromosome (Vicoso and Bachtrog 2013). The only sex-limited sex chromosome with sufficient size and annotation was the *S mansoni* W, which is also included.

We tested the significance of the over- or underrepresentation of mitochondrial genes on the sex chromosomes by bootstrapping. To calculate confidence intervals for sex chromosome bias, for each species/clade we sampled with replacement 10,000 times the number of genes on the sex chromosome, summed the number of genes with mitochondrial annotation, calculated bias (as above) and took the 95% confidence intervals of the distribution. To calculate confidence intervals for the autosomes bias, we sampled with replacement 1000 times the genes on each of the autosomes (i.e. between 4 and 27 chromosomes, depending upon the clade), calculated bias for each chromosome, calculated the mean bias for each sampling event, and calculated the 95% confidence intervals of the mean (i.e. the CI was calculated from 1000 samples, and each sample was the mean bias of all chromosomes). For each analysis, we corrected for multiple testing for 9 different sex chromosomes, at an alpha of 0.05, using Bonferroni correction (P < 0.0057). Sex chromosomes had a significant over- or underrepresentation of mitochondrial genes if the sex chromosome confidence interval did not overlap the confidence interval of the autosomes.

When grouping different species together (the Theria, as well as Aves) or when one species has multiple sex chromosomes (*O. anatinus*), we calculated the confidence interval for sex chromosome bias by summing together all the genes on the sex chromosomes and treating them as one large sex chromosome. When testing the autosomal distribution of the grouped species, sampling with replacement was done from each species such that each species contributed equally to the sampling distribution (i.e. to the 1000 bootstrapped data points). We tested whether the bias of neo-, ancient- and sex-limited chromosomes was different from the autosomes by bootstrapping all autosomal genes and excluding the homogametic sex chromosome.

We tested the significance of the overall over- or underrepresentation of mito-nuclear genes on the sex chromosomes in male- and female-heterogametic systems by bootstrapping 10,000 times the bias for each orthologous sex chromosome (mean bias for those sex chromosomes represented by multiple species) and calculating the 95% confidence intervals for X and Z chromosomes. This slightly different approach to the previous bootstrapping technique enabled each clade to contribute equally to the distribution, irrespective of the size of the sex chromosome.

The significance of over- or underrepresentations of mito-nuclear genes on the sex chromosomes were also analysed using Chi-Squared Tests.

It has been previously suggested that the paucity of mito-nuclear genes on the therian X chromosome was driven by sexual conflict related to asymmetrical inheritance (Drown, et al. 2012). Mito-nuclear genes have been suggested to move off the X onto autosomes due to conflict between the sexes (Drown, et al. 2012). This process would involve gene duplication and fixation, followed by loss of the sex-chromosome linked parent copy (Drown, et al. 2012; Gallach, et al. 2010). Genes with effects that can ameliorate male-detrimental mitochondrial mutations would be selected in males and are more likely to accumulate on autosomes than on female-biased X chromosomes. Although some have suggested that there should be a random distribution of mito-nuclear genes on Z chromosomes (Drown, et al. 2012), others have predicted an over-representation of mito-nuclear genes on the Z chromosome of female heterogametic species related to sexual selection (Hill and Johnson 2013).

These data indicate that, in addition to previous work on the therian mammals, only *C. elegans* also shows a non-random distribution of mito-nuclear genes on the X chromosome. Like the pattern in Theria, the *C. elegans* X shows an underrepresentation. The majority of the species studied here, both male and female-heterogametic, show no significant overall bias in their distribution of mito-nuclear genes. This suggests that patterns of mito-nuclear gene distribution are not shaped by convergence of sexual conflict over asymmetrical inheritance across independent sex chromosome systems. This pattern was consistent across all rBBH approaches (**Figure 4.1, Figure 4.2; Table 4.1, Table 4.2**) and species-specific GO annotations (**Figure 4.3, Table 4.3**).

We also explored the neo-X chromosome in *G. aculeatus* (Kitano, et al. 2009; Natri, et al. 2013) and the B chromosome in *D. melanogaster*, which has recently been shown to be an ancient sex chromosome that has reverted to an autosome in the *Drosophila* lineage (Vicoso and Bachtrog 2013), in order to test whether recent and past evolutionary history shaped current patterns.

Both the *G. aculeatus* X and neo-X showed no significant bias of mito-nuclear genes (**Table 4.1, Table 4.2, Table 4.3**). The ancient X chromosome in *D. melanogaster* also showed no overall bias (**Table 4.1, Table 4.2**). Finally, we also examined the W chromosome in *S. mansoni*, which is sufficiently large for such an analysis. However, no significant bias of mito-nuclear genes on the W was found (**Table 4.1, Table 4.2)**.

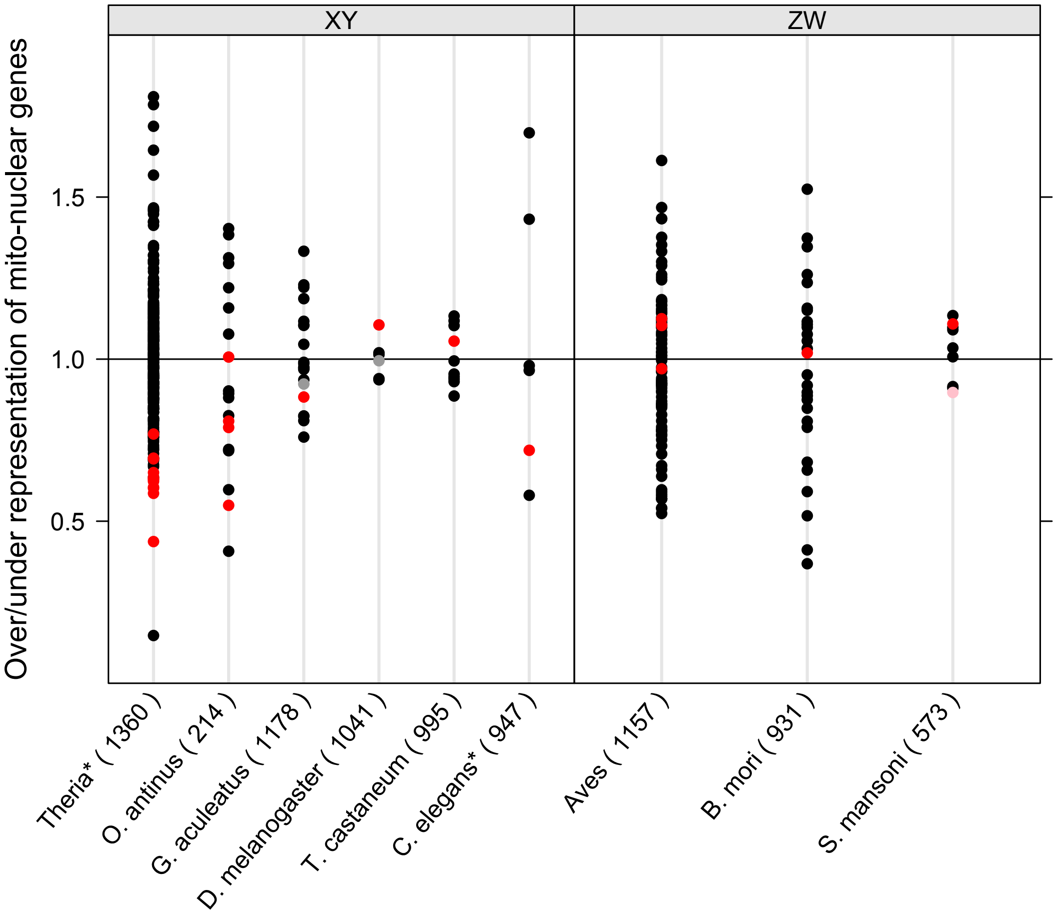


Figure 4.1 Bias of nuclear-mitochondrial genes on the sex chromosomes across species with independent sex chromosomes. Values for each autosome are in black, major sex chromosomes (X or Z) in red, old (i.e. *D. melanogaster* 4th) and neo (i.e. *G. aculeatus* Chromosome 9) in grey, and the *S. mansoni* W chromosome in pink. Values in parentheses after species name indicate the total number of mito-nuclear genes in the genome detected by the rBBH analysis with *M. musculus*. Species marked by * have a significant underrepresentation of nuclear-mitochondrial genes on the X chromosome. Note: Some of *D. melanogaster* autosomal points overlap.

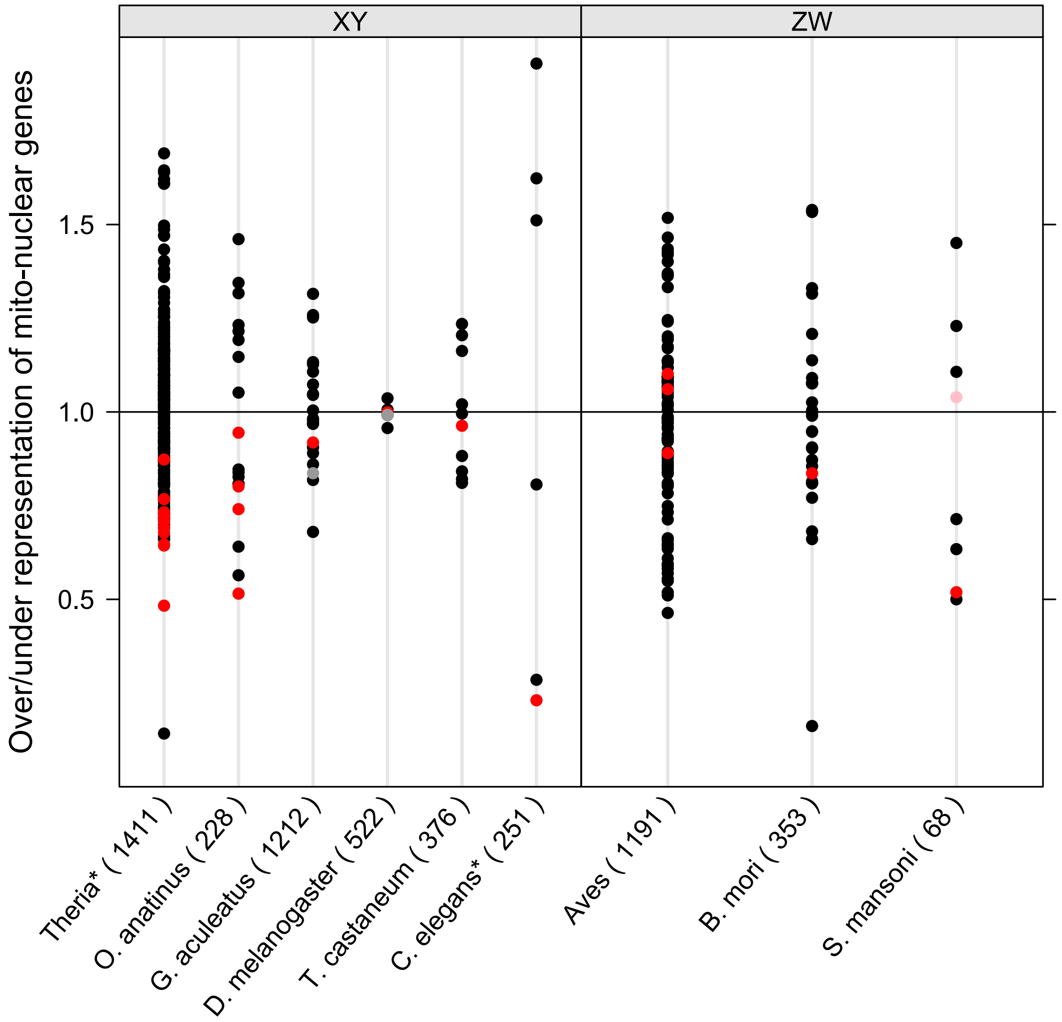


Figure 4.2 Bias of nuclear-mitochondrial genes on the sex chromosomes across species with independent sex chromosomes. Values for each autosome are in black, major sex chromosomes (X or Z) in red, old (i.e. *D. melanogaster* 4th) and neo (i.e. *G. aculeatus* Chromosome 9) in grey, and the *S. mansoni* W chromosome in pink. Values in parentheses after species name indicate the total number of mito-nuclear genes in the genome detected by the rBBH analysis using *M. musculus, D. melanogaster* and *C. elegans* to find orthologs. Species marked by * have a significant underrepresentation of nuclear-mitochondrial genes on the X chromosome. Note: Some of *D. melanogaster* autosomal points overlap.

Table 4.1 Mean bias and 95% confidence intervals of mito-nuclear genes on the sex chromosomes and autosomes. Significant under or overrepresentations are in bold. Confidence intervals calculated by bootstrapping. Chi-Squared statistics are also presented. 1:1 Orthologs were identified using *M. musculus* as the reference.

Species or clade	**Over/under representation of mito-nuclear genes on sex chromosome (bias)**	**95% Bonferroni-corrected CI of the sex chromosome**	**95% Bonferroni-corrected CI of the autosomes**	**Chi-Square Test and *P*-value**
Male heterogamety	0.86	0.72-1.00		
Therian mammals	Under (mean=0.64)	0.55-0.72	0.90-1.13	89.5, *P*<**0.0001**
H. sapiens	0.63			
P. troglodytes	0.69			
G. gorilla	0.62			
P. abelii	0.60			
M. mulatta	0.65			
E. caballus	0.59			
B. taurus	0.64			
S. scrofa	0.77			
O. cuniculus	0.63			
R. norvegicus	0.60			
M. musculus	0.69			
M. domestica	0.44			
O. anatinus	Under (mean=0.85)	0.45-1.26	0.64-1.27	0.92, *P*=0.34
G. aculeatus	Under (0.88)	0.57-1.20	0.92-1.09	0.93, *P*=0.33
D. melanogaster	Over (1.11)	0.89-1.33	0.77-1.23	2.17, *P*=0.14
T. castaneum	Over (1.06)	0.69-1.42	0.91-1.11	0.18, *P*=0.67
C. elegans	Under (0.72)	0.51-0.92	0.98-1.18	12.06, *P*=**0.0005**
Female heterogamety	1.06	1.02-1.11		
Aves	Over (mean=1.07)	0.86-1.28	0.86-1.09	0.92, *P*=0.34
G. gallus	1.10			
M. gallopavo	0.97			
T. guttata	1.12			
B. mori	Over (1.02)	0.61-1.43	0.86-1.04	0.01, *P*=0.90
S. mansoni	Over (1.11)	0.61-1.60	0.87-1.17	0.41, *P*=0.52
Sex-limited /neo/ancient				
G. aculeatus neo-X	Under (0.92)	0.57-1.19	0.92-1.09	0.47, *P*=0.59
D. melanogaster ancient-X (chromosome 4)	1.00	-0.08-2.08	0.91-1.09	0.00, *P*=0.97
S. mansoni W	Under (0.90)	0.63-1.16	0.85-1.18	1.00, *P*=0.32

Significant *P*-values are shown in bold

Table 4.2 Mean bias and 95% confidence intervals of mito-nuclear genes on the sex chromosomes and autosomes. Significant under or over-representations are in bold. Confidence intervals calculated by bootstrapping. Mito-nuclear genes detected by the rBBH analysis using *M. musculus, D. melanogaster* and *C. elegans* to find orthologs.

Species or clade	Over/under representation of mito-nuclear genes on sex chromosome (bias)	95% Bonferroni-corrected CI of the sex chromosome	95% Bonferroni-corrected CI of the autosomes	Chi-Square Test and *P*-value
Male heterogamety				
Therian mammals	Under (mean=0.71)	0.61-0.79	0.90-1.13	62.8, *P*<**0.0001**
H. sapiens	0.73			
P. troglodytes	0.69			
G. gorilla	0.72			
P. abelii	0.69			
M. mulatta	0.72			
E. caballus	0.64			
B. taurus	0.71			
S. scrofa	0.87			
O. cuniculus	0.77			
R. norvegicus	0.65			
M. musculus	0.68			
M. domestica	0.48			
O. anatinus	Under (mean=0.83)	0.43-1.22	0.69-1.29	1.38, *P*=0.24
G. aculeatus	Under (0.92)	0.60-1.23	0.93-1.09	0.47, *P*=0.49
D. melanogaster	No bias (1.00)	0.70-1.30	0.86-1.13	0.00, *P*=0.99
T. castaneum	Under (0.96)	0.37-1.55	0.84-1.14	0.03, *P*=0.86
C. elegans	Under (0.23)	0.0-0.46	0.91-1.28	23.8, *P*<**0.0001**
Female heterogamety				
Aves	Over (mean=1.02)	0.83-1.22	0.86-1.09	0.10, *P*=0.75
G. gallus	1.06			
M. gallopavo	0.89			
T. guttata	1.10			
B. mori	Under (0.84)	0.22-1.45	0.83-1.12	0.47, *P*=0.49
S. mansoni	Under (0.52)	-0.50-1.54	0.64-1.69	0.95, *P*=0.33
Sex-limited /neo/ancient				
G. aculeatus neo-X	Under (0.84)	0.54-1.13	0.92-1.09	1.96, *P*=0.16
D. melanogaster ancient-X (chr. 4)	Under (0.99)	-0.58-2.55	0.86-1.13	0.00, *P*=0.99
S. mansoni W	Under (1.04)	0.18-1.90	0.61-1.77	0.00, *P*=0.97

Significant *P*-values are shown in bold

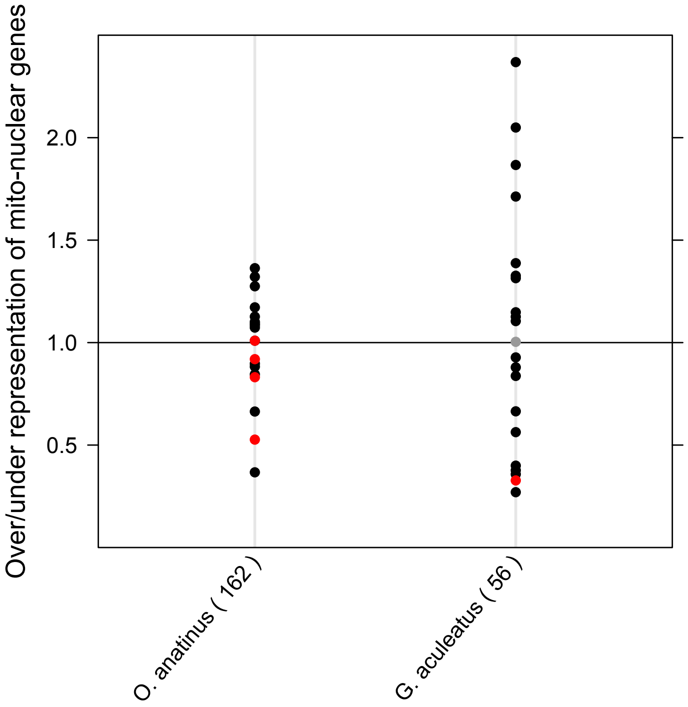


Figure 4.3 Bias of nuclear-mitochondrial genes on the sex chromosomes across *G. aculeatus* and *O. anatinus*. Values for each autosome are in black, X chromosomes in red, and neo (i.e. *G. aculeatus* Chromosome 9) in grey. Values in parentheses after species name indicate the total number of mito-nuclear genes in the genome detected using GO:0005739 to identify genes that interact with the mitochondria.

Table 4.3 Mean bias and 95 % confidence intervals of mito-nuclear genes on the sex chromosomes and autosomes. Mito-nuclear genes identified using Gene Ontology terms in Biomart.

Species or clade	Over/under representation of mito-nuclear genes on sex chromosome (bias)	95% Bonferroni-corrected CI of the sex chromosome	95% Bonferroni-corrected CI of the autosomes	Chi-Square Test and *P*-value
Male heterogamety				
O. anatinus	Under (mean=0.87)	0.41-1.33	0.36-1.35	0.60, *P*=0.44
G. aculeatus	Under (0.34)	-0.58-1.23	0.66-1.44	1.46, *P*=0.23
Sex-limited /neo/ancient				
G. aculeatus neo-X	1.00	-0.61-2.60	0.63-1.44	0.00, *P*=0.95

If sexual conflict over asymmetrical inheritance does shape the distribution of mito-nuclear genes, convergent patterns of underrepresentation across independent X chromosomes (Drown, et al. 2012) may be expected. X chromosomes have in general fewer mito-nuclear genes (i.e. bias < 1) than expected (mean bias = 0.86, CI = 0.72-1.00); however, only two of six independent X chromosomes showed statistically significant underrepresentations of mito-nuclear genes. The therian mammals exhibit the most extreme distribution of mito-nuclear genes on the X chromosome, with only the *C. elegans* X chromosome also showing a significant paucity. Furthermore, *C. elegans* is a gynodioecious species, with both males and hermaphrodites. The lack of distinct male and female individuals within the species may limit the degree of sexual conflict, as male-harming mutations in mito-nuclear genes would also affect the male function in hermaphrodites. This suggests that sexual conflict may be reduced in this species and may not be the driver of the distribution of mito-nuclear genes. However, it is important to note that gynodioecy is a recently derived trait in the *Caenorhabditis* lineage, and most other species in the genus are fully gonochoristic (each individual is either male or female). This means that any reduction in sexual conflict due to gynodioecy would have been relatively recent.

Many patterns driven by sexual conflict on X chromosomes are predicted to display converse patterns on Z chromosomes (Rice 1984), and this has been true of genomic characters, including the sexualisation of gene expression (Dean and Mank 2014). A convergent over-representation of mito-nuclear genes on Z chromosomes may be expected, although the low co-transmission between the mitochondria and the Z chromosome may ameliorate this prediction (Drown, et al. 2012). These results indicate that Z chromosomes overall have more mito-nuclear genes (i.e. bias > 1) than expected (mean bias = 1.06, CI = 1.02-1.11), but there was no taxon-specific case where a Z chromosome carried a significantly greater proportion of mito-nuclear genes than expected based on its relative size.

The W and mitochondria are in complete linkage, being co-transmitted each generation. Consequently, an over-representation of co-adapted, female-benefitting mito-nuclear genes on the W may be expected. Although we do observe some W-linked mito-nuclear genes in *S. mansoni*, suggesting that some genes have sex-specific expression, there is not an over-representation of these genes on the W. The lack of bias of mito-nuclear genes on the W could be due to lack of selection for gene movement in the female – the mitochondria is already optimised for females and so no advantage for the female is gained by movement of Z or autosomal genes onto the W.

It is possible that the genomic distribution of mito-nuclear genes is somewhat confounded by other genomic phenomena. First, mitochondrial mutation rate differs substantially across species; for example, mammals tend to have high rates and *Drosophila* have low rates (Montooth and Rand 2008). Mitochondrial mutation rate will affect the extent to which mitochondria can evolve female-beneficial mutations. Secondly, the relative rate of evolution of sex chromosomes to autosomes (the Faster-X Effect, Charlesworth, et al. 1987) varies across species, and depends upon the relative effective population size (Mank, et al. 2010). The relative effective population size of different X chromosomes to autosomes varies substantially (Mank, et al. (2010) and references therein); however, this does not necessarily explain these data, as, for example, *E. caballus* and *D. melanogaster* both have high relative effective population sizes of the X chromosome (Andolfatto 2001; Connallon 2007; Lau, et al. 2009; Singh, et al. 2007) and yet *D. melanogaster* shows no overall bias, while *E. caballus* shows an underrepresentation (**Table 4.1, Table 4.2**). Thirdly, we may expect variation in the magnitude of the male-biased mutation rate, for example due to species differences in generation time and in the strength of sexual selection and associated intensity of sperm competition (Ellegren 2007). However, it is difficult to see how the patterns we observe are driven by variation in male-biased mutation. Finally, levels of gene transfer and genome rearrangement are lineage-specific (Rand, et al. 2001), where low levels of movement will restrict the ability of different parts of the genome to respond to inter-genomic coadaptation and conflict. This may explain many of the non-significant associations.

Alternatively, interactions between the mitochondrial genome and the X and Z chromosome have been suggested to play a role in sexual selection and might be enriched for mito-nuclear genes that play a role in colouration, such as those involving carotenoids (Hill and Johnson 2013). We did not observe this predicted over-representation on any Z chromosomes, and it is difficult to see how differences among the study species in the degree and type of sexual selection explain the variance in the distribution of mitochondrial genes.

A further possibility is that the genomic distribution of mito-nuclear genes is driven by gametic function. Although mitochondrial activity is generally not crucial for non-motile egg function (de Paula, et al. 2013), it is integral to sperm energy production and motility (Cummins 2008). Although many genes are functionally diploid in sperm (Braun, et al. 1989), there is evidence that many genes are expressed within the spermatid and are subject to haploid selection (Joseph and Kirkpatrick 2004). Because any single spermatozoon will only carry either an X or Y chromosome, expression of mito-nuclear genes within the sperm would lead to selection against sex-linkage as half of the male gametes would lack a functional copy. Conversely, all sperm in female heterogametic species contain a Z chromosome, and there would be no expected selection against Z-linkage of mito-nuclear genes.

Furthermore, differences among taxa in sperm biology could explain some of the patterns we observe among male heterogametic taxa. For example, species differ in the presence or absence of sperm hyperactivation, which requires high mitochondrial activity (Cummins 2008). Also, the degree to which oxidative metabolism is required for sperm motility differs, and both human and mouse sperm do not need mitochondrial activity for motility (Cummins 2008). Factors such as this may affect the degree of haploid expression of mito-nuclear genes in sperm and therefore the distribution of mito-nuclear genes on X chromosomes. However, we hasten to point out that none of these explanations alone fully account for why Theria and *C. elegans* have an underrepresentation of mito-nuclear genes on their X chromosomes. More complex theory, taking into account patterns of gene duplication and gene movement, may be required to make sense of these patterns.

The need to maximize the number of independent sex chromosomes in these analyses means that some genomes with incomplete functional annotation were included. To solve this, we employed an rBBH approach in order to detect orthologs of mitochondrial interacting genes that are annotated in model organisms like *M. musculus*. However, this approach could be influenced by taxon-specific mito-nuclear genes and difficulties in orthology identification across large evolutionary distances. Although this does limit the number of genes we identify through strict orthology identification in some taxa, we do not believe that it has unduly biased our results for several reasons. First, nuclear genes that interact with the mitochondria are conserved across broad taxonomic groups (Lotz, et al. 2013; Porcelli, et al. 2007), suggesting that rBBH is broadly applicable. The convergence between the results using *M. musculus* as the reference for all rBBH with results using *D. melanogaster* and *C. elegans* as reference points recovered similar patterns, suggesting that conservation predominates over clade- or species-specific patterns. We also detected similar patterns using species-specific GO annotations.

In conclusion, our results are not universally consistent with either sexual conflict (Drown, et al. 2012) or sexual selection (Hill and Johnson 2013) driving the general distribution of mito-nuclear genes on all sex chromosomes. We observed significant underrepresentation of mito-nuclear genes in just two of six analysed X chromosomes, and no patterns of non-random distribution on any analysed Z chromosome. The results suggest that other genomic phenomena may limit the extent to which inter-genomic conflict (Drown, et al. 2012) or sexual selection (Hill and Johnson 2013) affect mito-nuclear distributions and confirm the importance of broad, phylogenetically independent analysis.

Chapter 5

Deficit of mito-nuclear genes on the human X chromosome predates sex chromosome formation

The analyses presented in this chapter have been published in *Genome Biology and Evolution*:

Dean R, **Zimmer F**, Mank JE (2015) Deficit of mito-nuclear genes on the human X chromosome predates sex chromosome formation. *Genome Biology and Evolution* 7(2):636-641, **doi**: 10.1093/gbe/evv017

Author contributions:

I designed the data analyses with Dr Rebecca Dean and wrote the paper in collaboration with Dr Rebecca Dean and Professor Judith Mank.

I contributed the bioinformatics analyses, including the synteny analysis and the development of a software to visualize the syntenic chromosomal regions. Dr Rebecca Dean and I analysed the gene movement.

Dr Rebecca Dean performed the statistical analyses to test for significant over- or underrepresentation of mitochondrial interacting genes on the X or Z chromosomes.

Summary

Two taxa studied to date, the therian mammals and *Caenorhaditis elegans*, display underrepresentations of mito-nuclear genes (mt-N genes, nuclear genes whose products are imported to and act within the mitochondria) on their X chromosomes. This pattern has been interpreted as the result of sexual conflict driving mt-N genes off of the X chromosome. However, studies in several other species have failed to detect a convergent biased distribution of sex-linked mt-N genes, leading to questions over the generality of the role of sexual conflict in shaping the distribution of mt-N genes. Here, we tested whether mt-N genes moved off of the therian X chromosome following sex chromosome formation, consistent with the role of sexual conflict, or whether the paucity of mt-N genes on the therian X is a chance result of an underrepresentation on the ancestral regions that formed the X chromosome. We used a synteny-based approach to identify the ancestral regions in the platypus and chicken genomes that later formed the therian X chromosome. We then quantified the movement of mt-N genes on and off of the X chromosome and the distribution of mt-N genes on the human X and ancestral X regions. We failed to find an excess of mt-N gene movement off of the X. The bias of mt-N genes on ancestral therian X chromosomes was also not significantly different from the biases on the human X. Together, these results suggest that, rather than conflict driving mt-N genes off of the mammalian X, random biases on chromosomes that formed the X chromosome could explain the paucity of mt-N genes in the therian lineage.

Introduction

A series of studies have recently generated substantial debate over the role of inter-genomic conflict in driving mito-nuclear gene distributions on and off sex chromosomes (Dean, et al. 2014; Drown, et al. 2012; Hill and Johnson 2013; Hough, et al. 2014; Rogell, et al. 2014). Mito-nuclear (mt-N) genes are loci whose products, encoded by the nuclear genome, are then imported into the mitochondria, which is the primary site of their activity. Because mitochondria and sex chromosomes have different inheritance patterns between the sexes, inter-genomic conflict has been suggested as a potential explanation for the underrepresentation of mt-N genes on the X chromosomes of some animals (Dean, et al. 2014; Drown, et al. 2012). Mitochondria are maternally inherited in many species (although low rates of male transmission may occur, e.g. Wolff, et al. 2013), and are therefore selected for female fitness effects, as male mitochondria are generally evolutionary irrelevant. It has been shown that maternal transmission of mitochondria can result in quite serious costs to males, through the disruption of male function (Drown, et al. 2012; Innocenti, et al. 2011; Partridge and Hurst 1998).

The accumulation of mutations that are detrimental to males could be ameliorated if genes that interact with the mitochondria move to a more favourable genomic location for the evolution of compensatory mechanisms. Genes on the X chromosome, which spend 2/3 of their time in females, are more often co-transmitted with mitochondria than autosomal genes (Rand, et al. 2001), and the X chromosome is also feminized in several species (reviewed in Dean and Mank 2014). This might make the X chromosome particularly unfavourable for male-biased compensation of the mitochondrial mutational load. It is therefore possible that there has been selection in males for the movement of mt-N genes off of the X chromosome in order to reduce disruption to male function induced by maternally -transmitted mitochondria.

Consistent with the conflict hypothesis, *Caenorhaditis elegans* (Dean, et al. 2014) and the therian mammals (Drown, et al. 2012) show a deficit of mt-N genes on their X chromosomes, and genes sensitive to mitochondrial polymorphism are scarce on the *Drosophila* X chromosome (Rogell, et al. 2014). However, a broader phylogenetic assessment of mt-N gene distributions revealed a mixed pattern, with most male heterogametic species studied showing no significant bias (Dean, et al. 2014; Hough, et al. 2014). Moreover, many sex-specific evolutionary properties observed on the X chromosome, such as distributions of sex-biased genes (Arunkumar, et al. 2009; Wright, et al. 2012) are observed in converse on Z chromosomes, so a corresponding overabundance of Z-linked mt-N genes in female heterogametic systems may be expected; however, no such overabundance has yet been observed (Dean, et al. 2014). Furthermore, if conflict is at least partly responsible for the genomic distribution of mt-N genes, it might also be expected to shape the distribution of nuclear genes that interact with the chloroplast, which is also often maternally inherited, but no bias was detected in the distribution of chloro-nuclear genes on the X chromosome in *Rumex* (Hough, et al. 2014), a dioecious plant with sex chromosomes.

These patterns of mt-N gene distributions suggest that either conflict is particularly strong only in therian mammals and nematodes, or that some effect other than conflict explains the distribution in these two clades. The incorporation of mitochondrial loci into the nuclear genome began long before the formation of sex chromosomes in any single extant lineage (Cortez, et al. 2014; Dyall, et al. 2004; Timmis, et al. 2004), and strong chromosomal biases exist for many autosomes, presumably due to chance variation in gene content (Dean, et al. 2014; Drown, et al. 2012; Hough, et al. 2014). This presents the possibility that biases in mt-N gene distributions need not be driven by conflict, but instead could predate the formation of the sex chromosome, if the precursor autosomes showed an ancestral bias through chance alone.

We tested whether ancestral gene distributions can explain the underrepresentation of mt-N genes on therian sex chromosomes. The rapid gene and genome evolution in *Caenorhaditis* (Lipinski, et al. 2011) precludes reconstruction of syntenic relationships across even closely-related species, but amniotes have conserved synteny (Dehal and Boore 2005), making it possible to identify syntenic regions in divergent taxa. In order to determine whether the paucity of mt-N genes on the therian X chromosome is a consequence of inter-genomic sexual conflict, or whether it is simply the product of a biased distribution on the ancestral autosome that gave rise to the therian X chromosome, we tested the mt-N gene distributions on the ancestral regions syntenic to the therian X in platypus and chicken (hereafter termed X-syntenic regions).

We used the human X chromosome as the point of reference because of its excellent annotation. Since the human X is broadly syntenic across therian mammals (Band, et al. 2000; Murphy, et al. 1999; Ohno 1967; Raudsepp, et al. 2004), it is representative of the therian X in general. We identified regions in synteny with the human X in platypus (*Ornithorhynchus anatinus*) and chicken (*Gallus gallus*), the most recent ancestors to the theria with different sex chromosomal systems (Graves 2006) and annotated genomes. This enabled us to use two complementary approaches to test the role of conflict in driving mt-N gene distributions. First, we identified orthologous genes, in platypus and chicken, to the human mt-N genes. We then tested for an excess of mt-N gene movement in order to investigate whether inter-genomic conflict has driven mt-N genes off of the human X following sex chromosome formation. Second, we used these orthologous genes to compare mt-N gene distributions on human X and X-syntenic regions in platypus and chicken. If the abundance of mt-N genes on the X-syntenic regions is more than the abundance on the human X, then inter-genomic conflict may have driven mt-N genes off of the therian X following sex chromosome formation. If, on the other hand, mt-N biases on the ancestral autosomes that gave rise to the therian X chromosome show a similar underrepresentation to the human X, then the chromosomal bias is unlikely to be a consequence of inter-genomic conflict and may simply be a result of random variation across chromosomes in mt-N content.

Materials and Methods

Identification of ancestral chromosomes to the human X chromosome

In the first step, I obtained the human (*Homo sapiens*), platypus (*Ornithorhynchus anatinus*) and chicken (*Gallus gallus*) proteomes from Ensembl version 76 (Flicek, et al. 2014). I used the longest isoforms as input for BLASTP (Altschul, et al. 1990) to detect homologs between the human proteome and both platypus and chicken (E-value < 10^{-10}). I then used the BLASTP output and positional information as input for MCScanX (Wang, et al. 2012), used with default values, to detect homologous chromosomal regions between human and platypus and human and chicken. Only genes that have been mapped to a chromosome were included for human and chicken; genes on UltraContigs were included for platypus, as a larger proportion of this genome assembly is currently mapped to scaffolds and contigs rather than chromosomes. The homologous chromosomal regions of the human X chromosome on platypus and chicken chromosomes were identified as the ancestral chromosomes to the human X chromosome. If the individual MCScanX alignments were closer than 10 million base pairs, I merged the alignments into a larger syntenic region to reflect the process of chromosome rearrangement (Burt, et al. 1999; Coghlan, et al. 2005) and sex chromosome formation (Lahn and Page 1999).

Identification of mt-N gene movement

Mt-N genes were identified in human using Gene Ontology annotation (GO:0005739) in Biomart Ensembl Genes 76 (Durinck, et al. 2005). To track movement of mt-N genes on and off the X, I identified one-to-one orthologs of the 1572 human mt-N genes in platypus and chicken using reciprocal best hit BLAST (rBBH), with a minimum E-value of 10^{-10}. Significant hits were ordered by bitscore and a rBBH was only counted when the tophit had a sequence identity larger than 30%. This resulted in 1064 rBBH between human and platypus, and 1116 between human and chicken. Of those, 575 rBBH between human and platypus, and 1087 between human and chicken, were on a sufficiently large scaffold to infer synteny (i.e. Ultra contigs in platypus and chromosomes in chicken).

To identify whether movement of mt-N genes on and off of the X chromosome represent an excess of gene movement, we calculated the expected number of movements based upon the number of genes on source chromosomes and the number of base pairs on the target chromosomes (Betrán, et al. 2002; Toups, et al. 2011; Vibranovski, et al. 2009). Fisher's Exact Tests were used to test whether observed movements were different from expected.

mt-N abundance

Gene counts of protein coding genes were calculated using Biomart Ensembl Genes 76. When comparing the abundance of mt-N genes on ancestral X and therian X between species, only the regions of the human X chromosome that were identified as syntenic in the other species were used. The bias of the distribution of mt-N genes on the human X and the platypus and chicken X-syntenic regions were calculated as: Bias = number of mt-N genes / expected number of mt-N genes, where the expected number was calculated as: Expected number = (number of genes in region / total genes) * total mt-N genes.

Mt-N genes in platypus and chicken were identified using two approaches; first, using the orthologous genes to the mt-N genes in human and second, using species–specific Gene Ontology annotation (GO:0005739) in Biomart Ensembl Genes 76. In chicken and platypus, GO:0005739 genes are inferred from electronic annotation (evidence code IEA), which includes sequence similarity, database records and keyword mapping files. As such, the orthology approach and the Biomart approach to infer gene function largely agree, with 76% overlap between the two approaches for platypus and 82% overlap for chicken.

Confidence intervals were calculated using 10,000 bootstrapped samples by randomly sampling genes with replacement and calculating the bias for each iteration. Differences between the expected and actual number of mt-N genes on the human X and platypus or chicken X-syntenic regions were calculated using a Fisher's Exact Test.

Results and Discussion

Mt-N gene movement on and off the human X chromosome

We identified platypus chromosome 6 plus ten unmapped ultra-contigs (platypus hX-syntenic regions), and regions of chicken chromosomes 1, 3, 4 and 12 (chicken hX-syntenic regions), as syntenic with the human X chromosome (**Figure 5.1**). The platypus hX-syntenic regions comprised a total of 381 genes spanning 71% of the length of the human X-chromosome and the chicken hX-syntenic regions comprised a total of 908 genes spanning 89% of the length of the human X-chromosome (**Figure 5.1**). The reduced coverage of the human X chromosome in platypus is largely due to the poorer assembly quality of the platypus genome.

To test whether an excess of mt-N gene movement off of the human X chromosome occurred following human X chromosome formation, we identified the location of the human mt-N orthologs in platypus and chicken. Pairs of orthologous genes that did not fall within syntenic blocks were potential candidates for genes that have moved. We identified four genes that moved onto the human X from Ultra contigs that were not in platypus hX-syntenic regions (from UltraContig 369; UltraContig 98; and two genes from UltraContig 519) and no genes that might have moved off the human X. These numbers were not significantly different from what is expected based on the relative size and content of the X chromosome (Betrán, et al. 2002; Toups, et al. 2011; Vibranovski, et al. 2009), (Fisher's Exact Test, *P* > 0.6), suggesting no excess of gene movement onto or off of the human X chromosome (**Table 5.1a**). However, two of the genes that might have moved onto the X were from Ultra Contig 519, part of which constitutes the platypus hX-syntenic region. Removing these genes does not qualitatively affect the results (Fisher's Exact Test, *P* > 0.2).

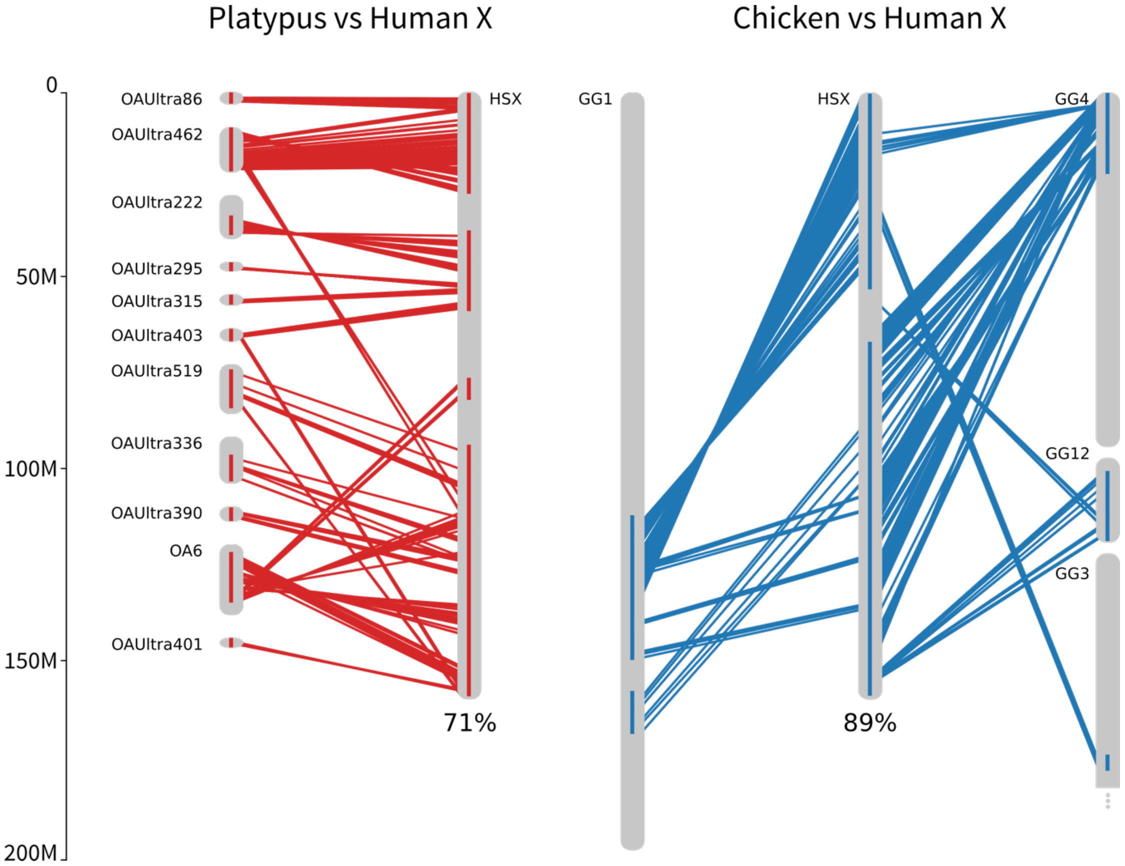


Figure 5.1 Syntenic regions between (a) human X (HSX) and platypus chromosome 6 (OA6) and several unmapped contigs (OAUltra) and (b) human X (HSX) and chicken chromosomes 1 (GG1), 4 (GG4), 3 (GG3) and 12 (GG12). Lines represent genes in synteny, red for platypus to human, blue for chicken to human. Blocks on chromosomes show regions where single MCScanX alignments are located on the chromosome closer than ten million base pairs.

Between human and chicken, three genes that moved onto the X (from GG8 and two from GG4) and three genes that moved off the X (to HS3 and two to HS2) were identified. This is not greater than expected based on the size of the X chromosome (Fisher's Exact Test, $P > 0.8$, **Table 5.1b**). Again, two of the genes that may have moved onto the X came from regions of GG4 that were close to the hX-syntenic region. These gene movements do not suggest an excess of mt-N gene movement off the human X (**Table 5.1b**, excluding two genes that might not have moved onto the X, Fisher's Exact Test $P > 0.3$). One of these genes (ENSP00000362773) was also found to move onto the X in platypus (platypus UltraContig 369 to HSX; chicken GG4 to HSX).

Table 5.1 Movement of mt-N genes on and off the X between (a) platypus and human and (b) chicken and human. X → A is hX-syntenic to autosome; A → X is autosome to human X syntenic region; A → A is autosome to autosome. *P*-value is from Fisher's Exact Test.

Movement	Observed	Expected[a]
(a) Platypus → human		
X → A	0	2
A → X	4	4
A → A	132	130
	P = 0.640	
(b) Chicken → human		
X → A	3	4
A → X	3	4
A → A	92	90
	P = 0.845	

[a]Calculated based on relative size and content of the X chromosome (Betrán, et al. 2002; Toups, et al. 2011; Vibranovski, et al. 2009)

mt-N gene abundance on X syntenic regions

The second approach was to compare the abundance of mt-N genes on human X chromosome regions that were syntenic to the identified regions in platypus and chicken. The bias (a measure of mt-N gene density, see methods) of mt-N genes does not differ between human X and platypus hX-syntenic regions (Fisher's Exact Test, *P* = 0.616**; Figure 5.2a, Table 5.2**) or human X and chicken hX-syntenic regions (Fisher's Exact Test, *P* = 0.793; **Figure 5.2, Table 5.2**), suggesting that the cause of the underrepresentation on the human X is more likely the result of a random underrepresentation of mt-N genes on the chromosomal regions that formed the human X, rather than inter-genomic conflict driving genes off of the X after its formation. We also calculated mt-N gene abundances using species-specific Gene Ontology annotation (GO:0005739) in Biomart to identify mt-N genes. The two approaches to infer mt-N gene function largely agree (platypus 76% overlap; chicken 82% overlap), hence calculating mt-N abundance using Biomart gave qualitatively similar results (**Figure 5.2b, Table 5.2**, human X and platypus hX-syntenic region, Fisher's Exact Test, *P* = 0.719; human X and chicken hX-syntenic regions, Fisher's Exact Test, *P* = 0.893).

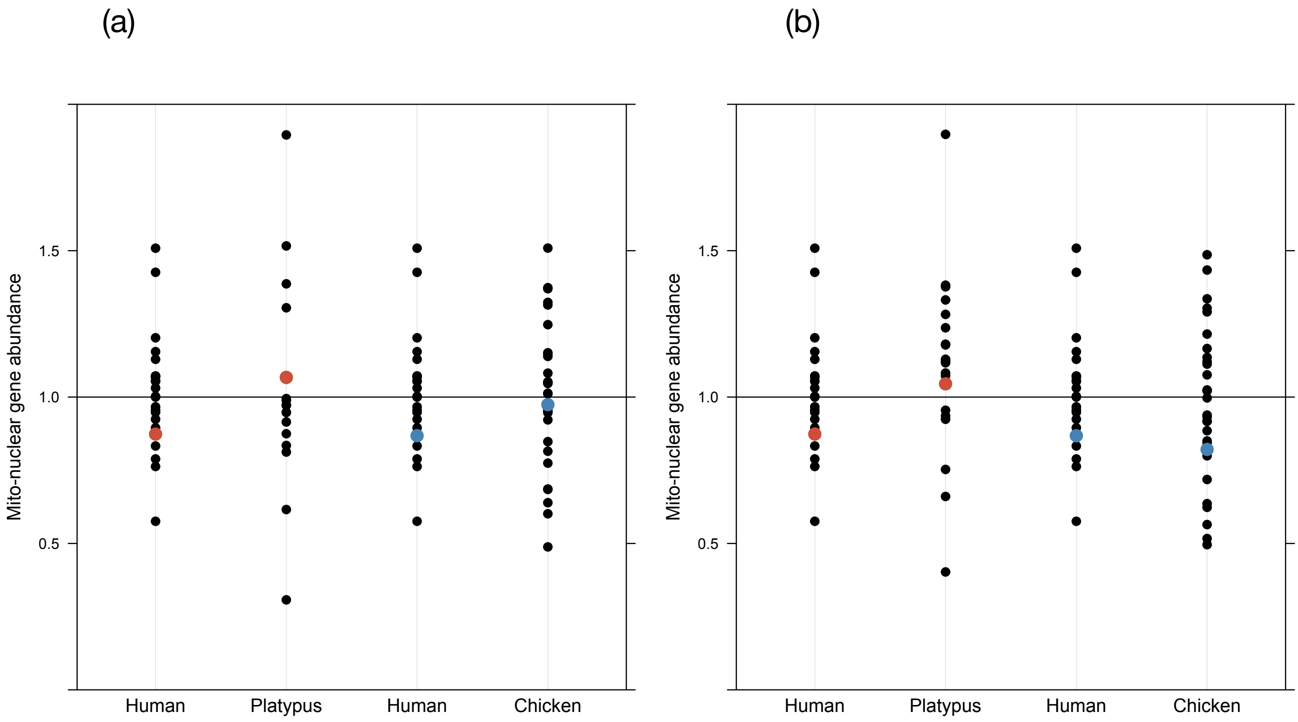


Figure 5.2 Bias of mt-N genes in human, platypus and chicken. Autosomes in black and hX-syntenic regions with platypus in red, hX-syntenic regions with chicken in blue. (a) Mt-N genes are inferred using orthology with human mt-N genes, and total gene counts include only those genes that are orthologous between human and platypus or human and chicken. (b) Mt-N genes are inferred through species-specific annotations in Biomart and gene counts are all annotated genes.

Table 5.2 Number of mt-N, total number of genes and the bias in distribution of mt-N genes on the human X and X-syntenic regions using gene orthology to identify mt-N genes and using species-specific mt-N gene annotations in Biomart. Note: gene counts are for the hX-syntenic blocks, the boundaries of which are created by merging alignments when alignments were closer than 10 million base pairs. The greater number of orthologous genes on chicken hX-syntenic than on the human X syntenic with chicken region is a consequence of these merged alignments.

Species	mt-N genes	Total genes	Bias	95% CI
Human X	55	820	0.85	0.64-1.06
Platypus hX-syntenic (orthology)	29	309	1.07	0.70-1.43
Platypus hX-syntenic (Biomart)	23	381	1.05	0.63-1.45
Human X (syntenic platypus)	46	667	0.87	0.63-1.12
Chicken hX-syntenic (orthology)	64	727	0.97	0.75-1.20
Chicken hX-syntenic (Biomart)	52	908	0.83	0.60-1.04
Human X (syntenic chicken)	49	715	0.87	0.64-1.10

Gene annotation and mt-N abundance

The measure of abundance (bias) relies on the total number of mt-N genes and total number of genes annotated in each species. This means measures of bias are susceptible to variation in the quality of genome annotation. The underrepresentation of mt-N genes on the whole of the human X in this study is 0.86 ± 0.22 (bias ± 95% CI), which is less pronounced than the underrepresentation previously reported for the human X chromosome (Dean, et al. 2014; Drown, et al. 2012). The human genome assembly version has recently been updated from GrCH37 to GrCH38, resulting in changes to the total number of genes and number of mt-N genes which can account for the different mt-N bias on the human X (bias ± 95% CI, 0.76 ± 0.21 using GrCH37). Gene annotation quality also likely accounts for the over-abundance of mt-N genes on the platypus hX-syntenic regions (29 observed mt-N genes and 25 expected), despite a lack of mt-N gene movement off of the X chromosome following X chromosome formation.

Mt-N gene abundance across independent X chromosomes

Across the seven independent X chromosomes studied to date, two (therian mammals and *C. elegans*) show a significant underrepresentation of mt-N genes, three (*Rumex*, platypus and stickleback) exhibit a non-significant underrepresentation, and two (*Tribolium* and *Drosophila*) show a non-significant over-representation (Dean, et al. 2014; Drown, et al. 2012; Hough, et al. 2014). This does not represent a significant overall underrepresentation of mt-N genes on X chromosomes (Two-tailed Sign-Test; 5 of 7, $P = 0.453$). If the distribution of mt-N genes on X chromosomes is explained by variation in ancestral autosomes, both under- and overrepresentations of mt-N genes on X chromosomes would be expected. This is consistent with what we find; however, the ability to detect a significant widespread underrepresentation (i.e. the signature of conflict) is not particularly powerful, with only 7 different X chromosomes having been quantified so far. An alternative explanation is that mt-N interactions predispose chromosomes depauperate of mt-N genes to become sex chromosomes, although this predisposition might be rather weak and highly dependent upon the location of genes involved in sex determination.

Taken together, these results suggest that the underrepresentation of mt-N genes on the therian X is not a result of gene movement off of the X chromosome. Rather, the paucity of mt-N genes on the therian X predates the formation of the therian sex chromosomes, and selection has acted mainly to maintain this ancestral distribution after sex chromosome formation. Even though we find no support for conflict driving mt-N genes off the therian X chromosome, random genomic biases in mt-N gene distributions could have important consequences for mt-N co-adaptation and potentially for sex chromosome formation. A paucity of mt-N genes on the therian X chromosome means that genes that interact with the mitochondria are less often co-transmitted compared to mt-N genes on autosomes. This might affect rates of co-evolution between mitochondria and nuclear genes (e.g.Hill 2014), with possible fitness consequences (Meiklejohn, et al. 2013; Montooth, et al. 2010).

Chapter 6

Phylogenetic analysis supports a link between DUF1220 domain number and primate brain expansion

The analyses presented in this paper have been published in *Genome Biology and Evolution*:

Zimmer F, Montgomery SH 2015. Phylogenetic analysis supports a link between DUF1220 domain number and primate brain expansion. *Genome Biology and Evolution* 7(8): 2083-2088, **doi**: 10.1093/gbe/evv122

Author contributions:

I designed the data collection, data analyses and wrote the paper in collaboration with Dr Stephen Montgomery.

I contributed all bioinformatics analyses, including the counting of the DUF1220 domains and evaluation of the variation in DUF1220 domain content. I also created all figures and wrote all programming code.

Dr Stephen Montgomery performed the phylogenetic gene-phenotype analyses and provided data on primate brain volumes.

Summary

In this chapter, we explore the phenotypic relevance of protein domain duplications using comparative genomic and phenotypic data from 12 primates. The expansion of DUF1220 domain copy number during human evolution is a dramatic example of rapid and repeated domain duplication. Although patterns of expression, homology and disease associations suggest a role in cortical development, this hypothesis has not been robustly tested using phylogenetic methods. Here, we estimate DUF1220 domain counts across 12 primate genomes using a nucleotide Hidden Markov Model. We then test a series of hypotheses designed to examine the potential evolutionary significance of DUF1220 copy number expansion. Our results suggest a robust association with brain size, and more specifically neocortex volume. In contradiction to previous hypotheses, we find a strong association with postnatal brain development, but not with prenatal brain development. Our results provide further evidence of a conserved association between specific loci and brain size across primates, suggesting human brain evolution may have occurred through a continuation of existing processes.

Introduction

The duplication of genes or chromosomes increases the number of copies present in the genome, thereby increasing gene dose. The concept of gene dose can be extended to protein domains, evolutionary conserved motifs that form three-dimensional units with distinct functions (Buljan and Bateman 2009). Most genes harbour more than one protein domain, resulting in specific domain arrangements (Chothia, et al. 2003). In comparison to the number of known protein domains, the number of different arrangements of these domains is much larger and is a driver of protein complexity (Levitt 2009; Vogel and Chothia 2006). Some protein domains undergo a rapid increase in copy number, similar to tandem repeats of gene duplicates (Björklund, et al. 2006), which can be seen as an increase in domain dose on a sub-gene level. The repeated addition of domains to a single gene may have a significant impact on the function of that gene. Charting the evolution of domain duplication provides an opportunity to investigate the phenotypic relevance of this effect.

The increase in DUF1220 domains during human evolution provides one of the most dramatic increases in copy number characterized in published genomes (Dumas, et al. 2012; Popesco, et al. 2006). A single copy of this protein domain is found in *PDE4DIP* in most mammalian genomes. In primates, this ancestral domain has been duplicated many times over, reaching its peak abundance in humans, where several hundred DUF1220 domains exist across 20-30 genes in the Nuclear Blastoma Breakpoint Family (NBPF) (Dumas, et al. 2012; Vandepoele, et al. 2005). The majority of these NBPF genes map to 1q21.1, a chromosomal region with complex, and unstable genomic architecture (O'Bleness, et al. 2014; O'Bleness, et al. 2012).

Interspecific variation in DUF1220 count show a pattern of phylogenetic decay with increasing distance from humans (Dumas and Sikela 2009; Dumas, et al. 2012; Popesco, et al. 2006). In humans, DUF1220 dosage has been linked to head circumference (Dumas, et al. 2012), and severe neurodevelopmental disorders, including autism spectrum disorder (ASD) and microcephaly (Davis, et al. 2014; Dumas, et al. 2012). The severity of ASD impairments is also correlated with 1q21.1 DUF1220 copy number, suggesting a dosage effect (Davis, et al. 2014). Taken together, these observations demonstrate how variation in domain content can have functional effects and have led to the suggestion that the expansion of DUF1220 copy number played an important role in human brain evolution (Dumas and Sikela 2009; Keeney, et al. 2014).

Although functional data is limited, they provide some indication on how DUF1220 domain copy number could influence brain development. DUF1220 domains are highly expressed during periods of cortical neurogenesis, suggesting a potential role in prolonging the proliferation of neural progenitors by regulating centriole and microtubule dynamics to control key cell fate switches critical for neurogenesis (Keeney, et al. 2015a). *PDE4DIP*, which contains the ancestral DUF1220 domain, does indeed associate with the spindle poles (Popesco, et al. 2006) and is homologous to *CDK5RAP2*, a centrosomal protein essential for neural proliferation (Bond, et al. 2005; Buchman, et al. 2010), which co-evolved with brain mass across primates (Montgomery, et al. 2011).

Two previous analyses report a significant association between DUF1220 copy number and brain mass, cortical neuron number (Dumas, et al. 2012), cortical grey and white matter, surface area and gyrification (Keeney, et al. 2015a). However, several limitations in these analyses restrict confidence in the results. First, DUF1220 copy number was assessed across species using a BLAT/BLAST analysis with a query sequence from humans, which introduces a bias that could partly explain the observed phylogenetic decay. Secondly, counts were not restricted to those domains occurring in functional exonic sequence and therefore many DUF1220 domains found in human pseudogenes were included in the analyses. Thirdly, the analyses were limited to a small number of species (4-8 species of primate), using parametric statistics that may not be suitable for count data and which do not correct for phylogenetic non-independence (Felsenstein 1985). This is not a negligible issue, because it can result in the overestimation of statistical significance (Carvalho, et al. 2006). Finally, previous phenotypic associations have been reported for multiple cortical phenotypes, all of which are strongly correlated with one another or are non-independent. Therefore, to date, these studies have not provided evidence for a specific association with neocortex size, nor have they tested the strength of the association with different periods of brain development which may provide new clues as to the functional relevance of DUF1220 domain copy number.

Here, we use nucleotide Hidden Markov Models implemented in HMMER3 (Eddy 2011; Wheeler and Eddy 2013) to more accurately query the DUF1220 domain number of distantly related genomes. After filtering these counts to limit the analysis to exonic sequence, we use comparative methods that correct for phylogenetic non-independence to test whether DUF1220 copy number is robustly associated with brain size, whether this is due to an association with pre- or postnatal brain development, and whether the association is specific to the neocortex.

Materials and Methods

Counting DUF1220 domains

HMMER3.1b (Eddy 2011) was used to build a Hidden Markov Model (HMM) from the DUF1220 (PF06758) seed alignment stored in the PFAM database (Finn, et al. 2014). The longest isoforms for all proteomes of 12 primate genomes from Ensembl v78 (Cunningham, et al. 2015) were searched using the protein DUF1220 HMM (hmmsearch, E-value < 10^{-10}) (**Table 6.1**). I extracted the corresponding cDNA regions to build a DUF1220 nucleotide profile HMM (nHMM), allowing for more sensitive analysis across a broad phylogenetic range. The DUF1220 nHMM was used to search the complete genomic DNA for all 12 species. These counts were filtered to remove any DUF1220 domains not located in annotated exonic sequence, or located in known pseudogenes.

I next filtered counts to limit them to exonic sequence in close proximity to the NBPF-specific Conserved-Mammal (CM) promoter (O'Bleness, et al. 2012). To do so, I built a nucleotide HMM for the CM promoter based on a MAFFT (Katoh, et al. 2002) alignment of the 900bp CM region upstream of human genes NBPF4, NBPF6 and NBPF7. Using this CM promoter nHMM, I searched 1000bp up- and downstream of genes containing DUF1220 domains for significant CM promoter hits (nhmmer, E-value < 10^{-10}). This provided final counts for DUF1220 domains within exonic regions and associated with the CM promoter (**Table 6.1**). These counts were used in subsequent phylogenetic analyses. All scripts and data used in the analysis are freely available from: https://github.com/qfma/duf1220

Phylogenetic gene-phenotype analysis

Phylogenetic Generalised Least Squares (PGLS) regressions were performed using *log*-transformed phenotypic data and *log*- or square root-transformed DUF1220 count data in BayesTraits (Pagel 1999). Phylogenetic multivariate generalized mixed models were implemented using a Bayesian approach in MCMCglmm (Hadfield 2010), to test for phylogenetically-corrected associations between DUF1220 counts and *log*-transformed phenotypic data (**Table 6.2**). All analyses were performed using a Poisson distribution, as recommended for count data (O'hara and Kotze 2010), with uninformative, parameter expanded priors for the random effect (G: V = 1,n v = 1, alpha.v = 0, alpha.V = 1000; R: V = 1, v = 0.002) and default priors for the fixed effects. Phylogenetic relationships were taken from the 10k Trees project (Arnold, et al. 2010). The posterior mean of the co-factor included in each model and its 95% confidence intervals (CI) is reported, and the probability that the parameter value is >0 (p_{MCMC}), as we specifically hypothesize a positive association (Dumas, et al. 2012). We explored the effects of alternative data treatments and transformations to test their influence on the reported phenotypic associations. Specifically, the analyses were repeated using Phylogenetic Least Square (PGLS) regressions using square-root and log_{10}-transformed DUF1220 counts to check whether or not the data transformation affected the relationships.

Two analyses were performed to test for heterogeneity and directional biases in the rate of change in DUF1220 counts across the primate phylogeny in Bayes Traits (Pagel 1999). To test for rate heterogeneity, we compared the fit of a one-rate Brownian-Motion (BM) model and a variable rates BM model to the data (Venditti, et al. 2011). These models are implemented using Markov chain Monte Carlo, run for 11,000,000 iterations with a burn-in of 1,000,000, and compared using *log*(Bayes Factors), calculated as 2(*log*[harmonic mean(complex variable rates model)] – *log*[harmonic mean(one rate model)]. The variable rates model accounts for rate heterogeneity by differentially stretching and compressing branch lengths and produces a posterior distribution of scaled phylogenies. The mean scaled branch lengths provide a visual indication of evolutionary rate when compared to the branch lengths of the input phylogeny. The posterior distribution of scaled phylogenies was subsequently used to test for a directional bias in DUF1220 domain variation by comparing the fit of a non-direction BM model to a directional BM model (Organ, et al. 2007; Pagel 1999).

Results

Variation in DUF1220 domain content

We used a Hidden Markov Model (HMM) based approach to detect DUF1220 in 12 primate genomes. First, we used a DUF1220 -based protein HMM to detect domains in annotated peptides. Previous DUF1220-based estimates were based on genomic data (Dumas, et al. 2012), and we used the initial hits based on the peptide HMM to construct a nucleotide HMM model (see Methods). Based on this nHMM, we find a varying number of DUF1220 domains in exonic sequences that are associated with the Conserved Mammal (CM) promotor and a large number of DUF1220 domains that are located in regions without any feature annotation (**Table 6.1**). Our genomic DUF1220 counts based on the nHMM model are largely similar to previous estimates (O'Bleness, et al. 2012) with some notable differences. In anthropoids, we find a significant negative relationship between the percentage deviation and time of divergence from *H. sapiens* ($P = 0.005$, $R^2 = 0.818$), suggesting the nucleotide HMM performs better than BLAT searches in more distantly related species. However, only ca. 20% of DUF1220 domains are found in exonic sequence data in most species (**Table 6.1**). The exception is *H. sapiens,* where almost all DUF1220 domains are in predicted coding sequence.

We explored the possibility that this difference may be due to annotation quality by estimating DUF1220 counts across different Ensembl versions for *H. sapiens*, *P. troglodytes* and *M. mulatta*. The number of DUF1220 domains is consistent across versions for both *P. troglodytes* and *M. mulatta*; in *H. sapiens*, however, there is a trend for DUF1220 counts to increase (**Figure 6.1**), most probably due to improvements in assembly and annotation (O'Bleness, et al. 2014). This either suggests that many DUF1220 domains in other primates occur in featureless regions of the genome, or that consistently poor annotation quality for *P. troglodytes* and *M. mulatta* leads to an underestimation of DUF1220 domains in exonic regions.

Table 6.1 DUF1220 count data across 12 primates using a protein HMM on peptide data and a nucleotide HMM (nHMM) for DNA data.

	O'Bleness *et al.* (2012)	**HMM**	**nHMM**				
Species	**whole genome**	**peptide**	**whole genome**	**exonic with CM**	**exonic with CM excl. pseudogenes**	**exonic without CM**	**no feature**
H. sapiens	272	246	302	298	262	0	4
P. troglodytes	125	37	138	34	34	9	95
G. gorilla	99	38	97	32	32	17	48
P. abelii	92	28	101	27	27	13	61
N. leucogenys	53	5	59	6	6	0	53
P. anubis	-	27	75	15	15	14	46
C. sabaeus	-	22	48	16	16	8	24
M. mulatta	35	21	74	10	10	13	51
C. jacchus	31	12	75	9	9	6	60
T. syrichta	-	2	47	2	2	1	44
M. murinus	2	0	4	1	1	1	2
O. garnettii	3	0	4	1	1	1	2

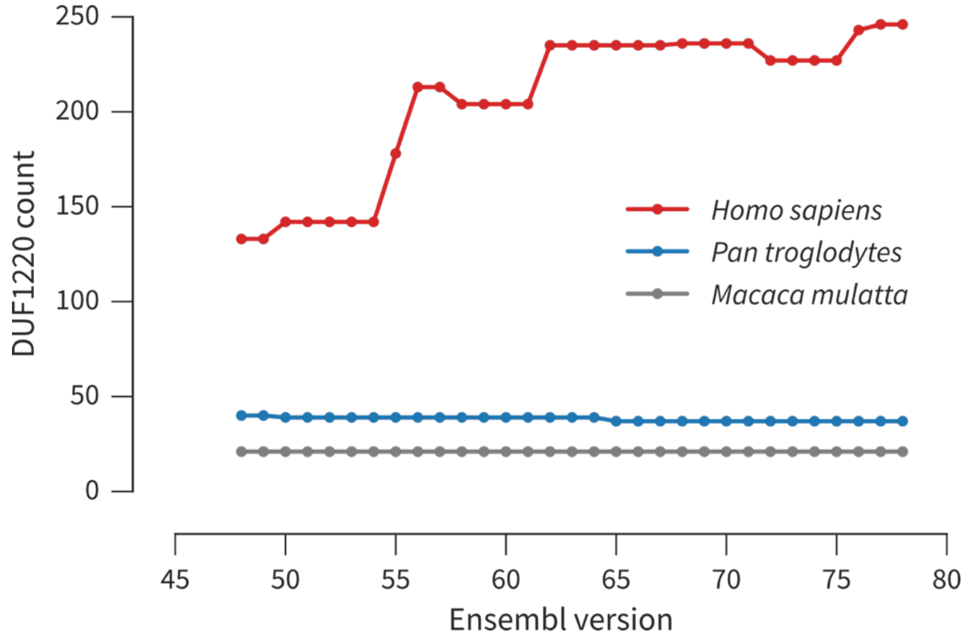


Figure 6.1 Stability of DUF1220 domain counts in peptides across different Ensembl versions

Exploring variation in evolutionary rate of DUF1220 domain number

We used counts of exonic DUF1220 domains associated with the CM promoter across all 12 primates to test for variations in the evolutionary rate of DUF1220 domain number. For these analyses we square-root transformed the DUF1220 counts. Square-root transformations may not adequately account for the large variance of count data, which is Poisson-distributed (O'hara and Kotze 2010), but see Ives (2015). This rates analysis should therefore be viewed with some caution, but provides an initial assessment of the phylogenetic patterns of DUF1220 domain number evolution. Using exonic DUF1220 domains associated with the CM promoter, we find evidence of heterogeneity in the rate of change in DUF1220 counts across the primate phylogeny (Bayes Factor = 10.254, 'very strong' support; **Figure 6.2**). We also find support for a directional model of expansion in DUF1220 number (Bayes Factor = 4.484, 'strong' support; **Figure 6.3**). A comparison of scaled branch lengths obtained from the variable rates model and the branch lengths of the input time-tree confirms that *H. sapiens* has the highest rate of increase (**Figure 6.4**). However, rates of evolution are high throughout hominoids (**Figure 6.4**), suggesting that DUF1220 increased independently in other lineages, and that the human expansion is an exaggeration of a general hominoid trend.

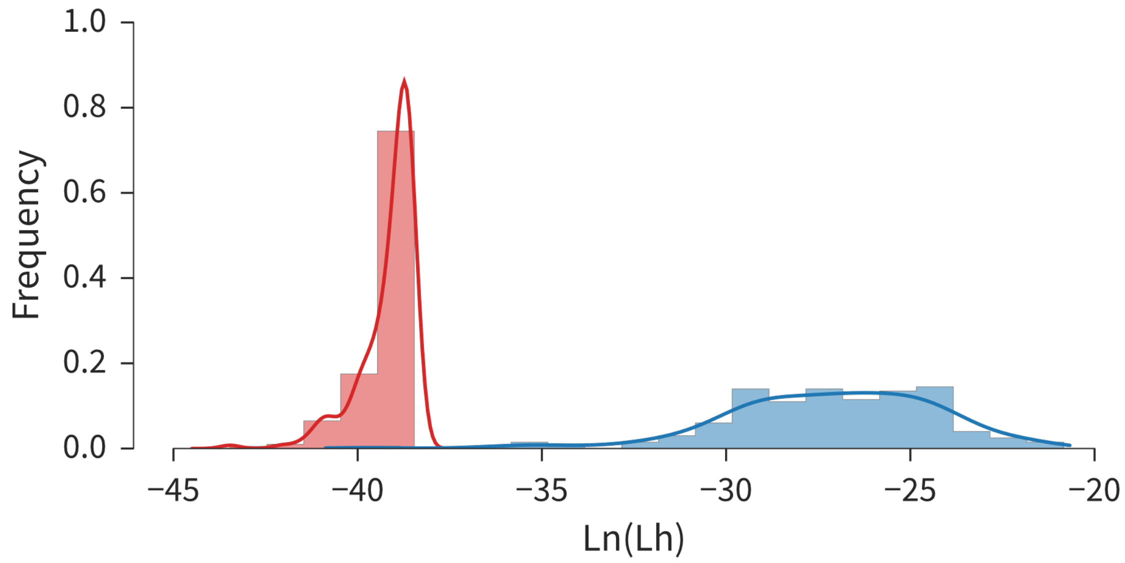


Figure 6.2 Distribution of likelihoods for one-rate (red) and variable rate (blue) models

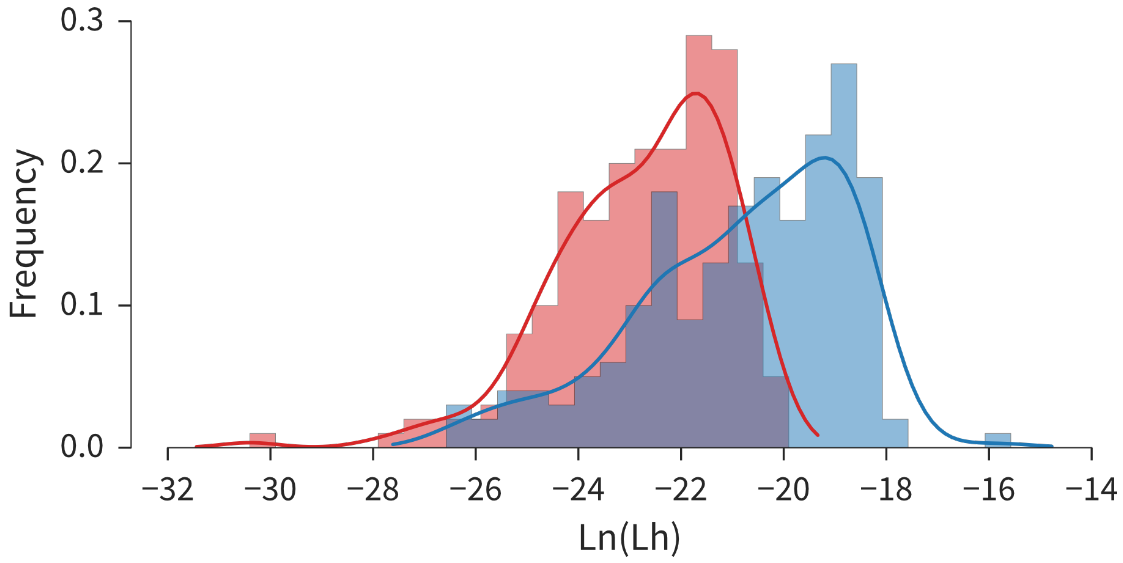


Figure 6.3 Distribution of likelihoods for non-directional (red) and directional (blue) models

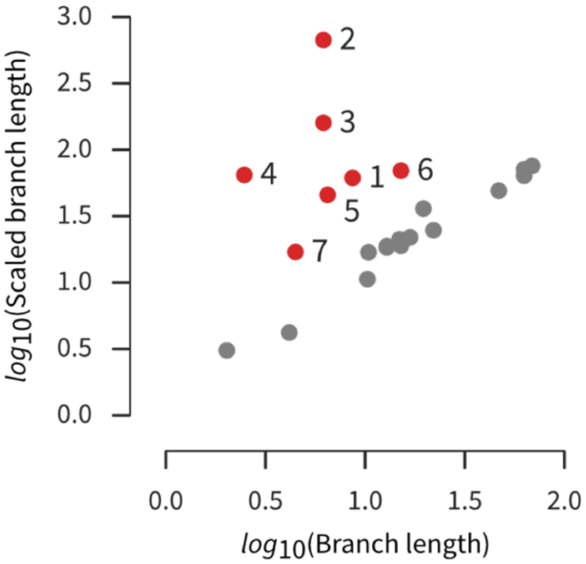


Figure 6.4 Plot of scaled-branch lengths from the variable rates model against the given branch length in time. Labelled nodes: 1) *G. gorilla* terminal branch, 2) *H. sapiens* terminal branch, 3) branch leading to the last common ancestor (LCA) of Hominini, 4) branch leading to the LCA of Homininae, 5) terminal *P. albelii* branch, 5) branch leading to the LCA of Hominidae. Branch lengths were *log*-transformed to compress the scale.

Gene-Phenotype co-evolution

Having confirmed significant inter-specific variation in DUF1220 counts across primates (**Table 6.1, Figure 6.2, Figure 6.3, Figure 6.4**), We next sought to test whether this variation co-evolves with phenotypic variation in brain size. The phenotypic data used for all analyses is shown in **Table 6.2.** We first used a Bayesian approach that corrects for phylogenetic non-independence and fits a Poisson distribution to the DUF1220 count data using MCMCglmm (Hadfield 2010). Additionally, we analysed the data using Phylogenetic Generalised Least Square (PGLS) regressions (Pagel 1999) using square-root transformed counts DUF1220, or log_{10}-transformed DUF1220. The square-root and log_{10} data transformations may not be appropriate for count data where models based on Poisson distributions provide more accurate results but are included to test how the association varies under different statistical assumptions (O’hara and Kotze 2010).

DUF1220 analysis using a Bayesian approach

Using a Bayesian approach that corrects for phylogenetic non-independence and fits a Poisson distribution to the DUF1220 count data using MCMCglmm (Hadfield 2010) we find evidence that CM-associated exonic DUF1220 counts are associated with brain mass across primates (n = 12, posterior mean =1.927, 95% CI = 0.800-3.040, $p_{MCMC} = 0.001$). This association is robust to the exclusion of *H. sapiens* (posterior mean =1.271, 95% CI = 0.490-2.019, $p_{MCMC} = 0.003$), and found when hominoids (n = 5, posterior mean = 3.679, 95% CI = 0.966-6.258, $p_{MCMC} = 0.018$) or anthropoids (n = 9, posterior mean = 2.019, 95% CI = 0.352-3.684, $p_{MCMC} = 0.010$) are analysed alone, suggesting a consistent phylogenetic association. When body mass is included as a co-factor in the model, the positive association is restricted to brain mass (**Table 6.3a, Figure 6.5a**).

Separation of pre- and postnatal development specifically links DUF1220 number to postnatal brain growth. Analysed separately, the association with prenatal brain growth is weaker (n = 11, posterior mean =1.758, 95% CI = -0.039-3.543, $p_{MCMC} = 0.023$) than with postnatal brain growth (posterior mean =1.839, 95% CI = 0.895-2.808, $p_{MCMC} = 0.001$). If both traits are included in the same model, only the positive association with postnatal brain growth remains (**Table 6.3b, Figure 6.5b**). Multiple regression analysis also confirms the association is specific to postnatal brain growth, rather than postnatal body growth (**Table 6.3b**).

Table 6.2 Phenotypic data used in the phylogenetic analyses

a) Adult brain volumes

Species[1]	**Brain mass (mg)**	**Body mass (g)**	**Brain volume (mm3)**	**Neocortex volume (mm3)**	**Cerebellum volume (mm3)**
H. sapiens	1330000	65000	1251847	1006525	137421
P. troglodytes	405000	46000	382103	291592	43663
G. gorilla	500000	105000	470359	341444	69249
P. abelii[2]	333000	54000	321429	219800	42900
N. leucogenys[3]	102000	5700	97505	65800	12078
M. mulatta	93000	7800	87896	63482	8965
P. anubis	201000	25000	190957	140142	18683
C. sabaeus[4]	108000	7800	103167	77141	8738
C. jacchus	7600	280	7241	4371	757
T. syrichta	3600	125	3393	1769	376
M. murinus	1780	54	1680	740	234
O. garnettii	10300	850	9668	4723	1414

[1] Unless otherwise indicated all data from Stephan, et al. (1981)

[2] Data from Zilles and Rehkämper (1988)

[3] Phenotypic data are from the closely related *Hylobates lar* and [4]*Erythrocebus patas*.

b) Pre and postnatal brain and body growth

	neonate		adult			
Species[1]	brain size (g)	body size (g)	brain size (g)	body size (g)	postnatal brain growth (g)	postnatal body growth (g)
H. sapiens	299.916	3318.945	1330.454	59292.532	1030.538	55973.588
P. troglodytes	153.109	1527.566	404.576	44977.985	251.467	43450.419
G. gorilla	289.068	2070.141	500.035	124738.351	210.967	122668.210
P. abelii	168.655	1603.245	364.754	56885.293	196.099	55282.048
N. leucogenys	65.013	354.813	102.094	5623.413	37.081	5268.600
M. mulatta	45.499	475.335	92.897	9908.319	47.398	9432.984
P. anubis[2]	77.268	831.764	160.325	14791.084	83.056	13959.320
C. sabaeus[3]	33.497	356.451	66.681	3732.502	33.184	3376.050
C. jacchus	3.631	28.184	7.603	358.922	3.972	330.738
T.syrichta[4]	2.748	25.177	3.771	119.437	1.024	94.260
M. murinus	-	-	-	-	-	-
O. garnettii	3.999	39.994	7.907	763.836	3.907	723.841

[1] Data from Capellini, et al. (2011).

[2] Phenotypic data was from the closely related species *P. hamadryas*, [3] *C. aethiops*, [4] mean of three con-generic

Table 6.3 MCMCglmm results of multivariate models

a) Brain mass and body mass

Model	**Posterior mean**	**95% CI**	p_{MCMC}
1. *log*(brain mass)	4.105	2.163 - 6.000	**0.001**
+ *log*(body mass)	-1.986	-3.544 - -3.900	0.988

b) Prenatal and postnatal growth

Model	**Posterior mean**	**95% CI**	p_{MCMC}
1. *log*(prenatal brain growth)	-2.158	-4.471 - 0.106	0.967
+ *log*(postnatal brain growth)	3.319	1.470 - 4.982	**0.002**
2. *log*(postnatal brain growth)	2.910	1.641 - 4.151	<**0.001**
+ *log*(postnatal body growth)	-1.241	-2.442 - -0.052	0.977

c) Brain regions

Model	**Posterior mean**	**95% CI**	p_{MCMC}
1. *log*(neocortex volume)	5.961	0.720 - 11.173	**0.014**
+ *log*(RoB volume)	-5.817	-13.322 - 1.120	0.953
2. *log*(cerebellum volume)	3.699	-5.857 - 12.611	0.186
+ *log*(RoB volume)	-2.435	-13.869 - 10.132	0.681
3. *log*(neocortex volume)	6.076	-0.139 - 12.5712	**0.025**
+ *log*(cerebellum volume)	-0.369	-9.5128 - 8.961	0.526
+ *log*(RoB volume)	-5.494	-15.814 - 5.288	0.872

Significant *P*-values are shown in bold

Finally, we examined the hypothesised relationship with neocortex volume (Keeney, et al. 2015a; Keeney, et al. 2014), but also consider cerebellum volume, as this region co-evolves with the neocortex (Barton and Harvey 2000), has expanded in hominoids (Barton and Venditti 2014), and shows high levels of NBPF expression (Popesco, et al. 2006). When the rest-of-the-brain (RoB) is included as a co-factor, to account for variation in overall brain size, a positive association is found for neocortex volume but not cerebellum volume (**Table 6.3c, Figure 6.5c**). To test whether the MCMCglmm results are robust to the effects of low genome quality on estimating DUF1220 domain counts, we repeated the key tests using total genomic DUF1220 counts (**Table 6.1**, 'whole genome'). The pattern of phenotypic associations is similar to those found with the exonic DUF1220 counts. There is evidence for a greater association with brain mass than body mass, with postnatal brain growth rather than prenatal brain growth and with neocortex volume **(Table 6.4**).

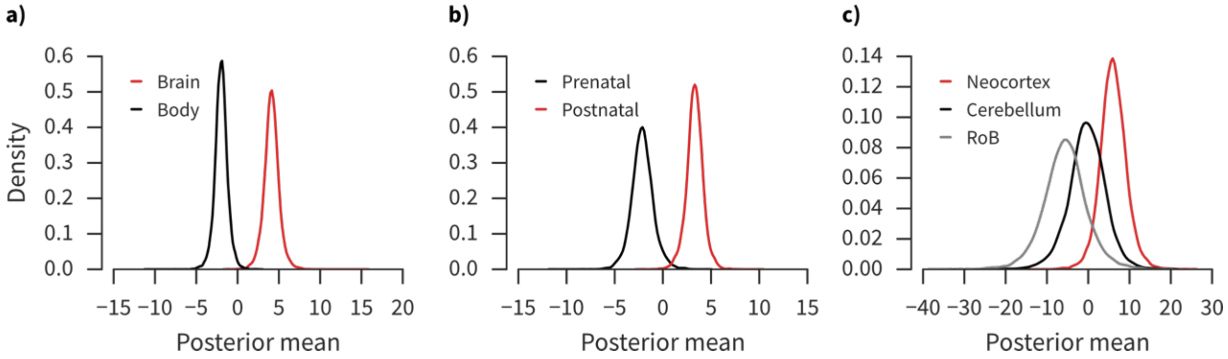


Figure 6.5 a) Posterior means of the association between DUF1220 count and brain mass (red) and body mass (black). b) Posterior means of the association between DUF1220 count and postnatal brain growth (red) and prenatal brain growth (black). c) Posterior means of the association between DUF1220 count and neocortex volume (red), cerebellum volume (solid black) and rest-of-brain volume (dashed black).

Table 6.4 MCMCglmm results using total genomic DUF1220 counts

Model	Posterior mean	95% C.I.	p_{MCMC}
1. Brain mass	2.130	0.249 – 4.039	0.014
+ body mass	-1.060	-2.589 - 0.479	0.929
2. Prenatal brain growth	-0.890	-3.198 – 1.383	0.815
+ postnatal brain growth	1.443	-0.336 – 3.060	0.043
3. Postnatal brain growth	1.432	0.2149 – 2.591	0.011
+ postnatal body growth	-0.691	-1.812 – 0.410	0.918
4. Neocortex volume	4.291	-1.089 – 9.3149	0.046
+ cerebellum volume	-1.7152	-8.975 – 5.501	0.710
+ rest-of-brain volume	-2.470	-10.915 – 6.039	0.741

Significant *P*-values are shown in bold

DUF1220 analysis using Phylogenetic Generalised Least Square (square-root transformed counts)

CM-associated exonic DUF1220 counts are significantly associated with brain mass across primates (t_{10} = 3.165, *P* = 0.005, R^2 = 0.455, **Figure 6.6a,b**), but are not associated with body mass (t_{10} = 0.922, *P* = 0.189, R^2 = 0.066). The relationship with brain mass is robust to the exclusion of *H. sapiens* (t_9 = 3.810, *P* = 0.002, R^2 = 0.569), and remains if body mass is included as a covariate in a multiple regression analysis (t_8 = 8.937, *P* < 0.001, R^2 = 0.878). The association with brain mass is also found when only anthropoids (t_7 = 3.196, *P* = 0.008, R^2 = 0.531) or only hominoids (t_5 = 4.976, *P* = 0.008, R^2 = 0.832), are included in the analysis, and when all hominoids are excluded from the analysis (t_7 = 2.749, *P* = 0.020, R^2 = 0.519).

Separation of pre- and postnatal development specifically links DUF12220 number to postnatal brain growth. Analysed separately, the association with prenatal brain growth is not significant (t_9 = 1.641, *P* = 0.067, R^2 = 0.197), but it is strongly significant for postnatal brain growth (t_9 = 3.850, *P* = 0.002, R^2 = 0.573). If both traits are analysed together in a multiple regression, the positive association with postnatal brain growth remains significant ($t_7 = 5.033$, $P < 0.001$, $R^2 = 0.757$), even if *H. sapiens* is excluded ($t_6 = 2.477$, $P = 0.021$, $R^2 = 0.466$), whilst prenatal brain growth is not ($t_7 = -2.879$, $P = 1.000$). Multiple regression analysis also confirms the association is specific to postnatal brain growth ($t_7 = 7.824$, $P < 0.001$, $R^2 = 0.853$), as opposed to body growth ($t_7 = -4.581$, $P = 1.000$).

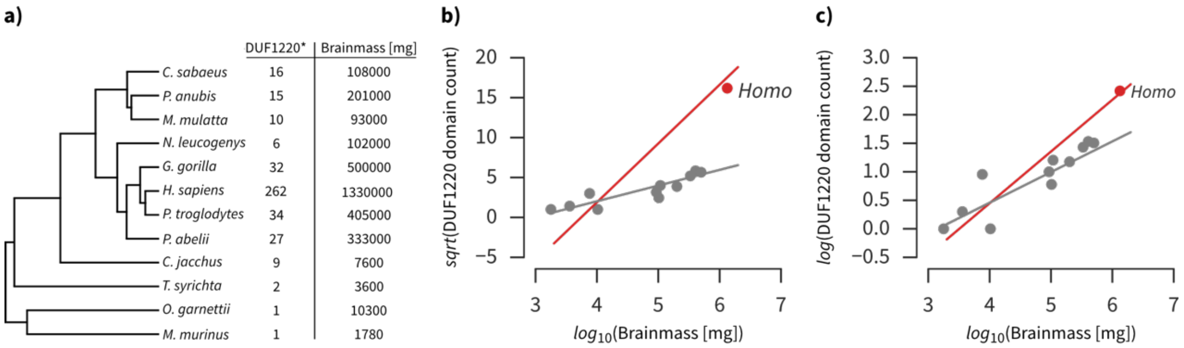


Figure 6.6 a) Phylogeny of Ensembl primates showing the number of DUF1220 domains in functional, annotated genes with a CM promoter, and brain mass. b) The relationship between square-root transformed DUF1220 counts and log_{10}(brain mass [mg]), and c) the relationship between log_{10} transformed DUF1220 counts and log_{10}(brain mass [mg]). The regression lines are shown with (red) and without (grey) the inclusion of the *H. sapiens* data. In all cases they are significant

The association is significant for neocortex volume (t10 = 3.162, *P* = 0.005, R2 = 0.454) and cerebellum volume (t10 = 2.767, *P* = 0.010, R2 = 0.390). However, when the volume of the rest-of-the-brain (RoB) is included as a covariate, the association with neocortex remains (t8 = 2.614, *P* = 0.015, R2 = 0.586; RoB: t8 = -1.952, *P* = 1.000), whilst the association with cerebellar volume is no longer significant (t8 = 1.220, *P* = 0.129, R2 = 0.421; RoB: t8 = -0.825, *P* = 1.000). A multiple regression between neocortex and cerebellar volume also suggests a stronger positive association with the neocortex (neocortex: t8 = 1.684, *P* = 0.065; cerebellum: t8 = -1.125, *P* = 1.00, R2 = 0.507).

DUF1220 analysis using Phylogenetic Generalised Least Square (*log*-transformed DUF1220 domain data)

CM-associated exonic DUF1220 counts are significantly associated with brain mass across primates (t10 = 4.770, *P* < 0.001, R2 = 0.655; **Figure 6.6a,c**). The association with body mass is weak (t10 = 1.880, *P* = 0.045, R2 = 0.228). The relationship with brain mass is robust to the exclusion of *H. sapiens* (t9 = 3.952, *P* = 0.002, R2 = 0.586), and remains if body mass is included as a covariate in a multiple regression analysis (t8 = 7.119, *P* < 0.001, R2 = 0.852). The association with brain mass is also found when only anthropoids (t7 = 4.100, *P* = 0.002, R2 = 0.651), or only hominoids (t5 = 10.165, *P* = 0.018, R2 = 0.954), are included in the analysis, and when all hominoids are excluded from the analysis (t7 = 2.455, *P* = 0.029, R2 = 0.462).

Separation of pre- and postnatal development specifically links DUF12220 number to postnatal brain growth. Analysed separately, the association with prenatal brain growth is significant (t9 = 2.435, *P* = 0.019, R2 = 0.351), but it is much more strongly significant for postnatal brain growth (t9 = 5.521, *P* < 0.001, R2 = 0.735). If both traits are analysed together in a multiple regression, the positive association with postnatal brain growth remains significant (t7 = 5.498, *P* < 0.001, R2 = 0.827), even if *H. sapiens* is excluded (t6 = 2.180, *P* = 0.032, R2 = 0.604), whilst prenatal brain growth is not significant (t7 = -2.419, *P* = 1.000). Multiple regression analysis also confirms the association is specific to postnatal brain growth (t7 = 7.564, *P* < 0.001, R2 = 0.862), as opposed to body growth (t7 = -3.197, *P* = 1.000).

The association is significant for neocortex volume (t10 = 4.869, *P* <0.001, R2 = 0.664) and cerebellum volume (t10 = 4.037, *P* = 0.001, R2 = 0.576). However, when the volume of the rest-of-the-brain (RoB) is included as a covariate, the association with neocortex remains (t8 = 3.525, *P* = 0.003, R2 = 0.775; RoB: t8 = -2.426, *P* = 1.000), whilst the association with cerebellar volume is no longer significant (t8 = 1.153, *P* = 0.141, R2 = 0.587; RoB: t8 = -0.563, *P* = 1.000). A multiple regression between neocortex and cerebellar volume also suggests a specific positive association with the neocortex (neocortex: t8 = 2.583, *P* = 0.016; cerebellum: t8 = -1.672, *P* = 1.000, R2 = 0.727).

Discussion

Changes in gene or domain dose are hypothesised to have a severe impact on the phenotype when dosage-sensitive genes or domains are affected. Our phylogenetic analyses substantiate the hypothesis that the increase in DUF1220 number co-evolves with brain mass (Dumas, et al. 2012; Keeney, et al. 2015a) and may contribute to the proximate basis of primate brain evolution. We have extended the results of previous studies by demonstrating specific associations with neocortex volume and postnatal brain growth, rather than prenatal brain growth. Together these results imply a role for DUF1220 in evolutionary changes in the maturation and postnatal development of the neocortex. Previous hypotheses concerning the phenotypic relevance of DUF1220 domain number have focused on their possible contribution to neurogenesis (Dumas and Sikela 2009; Keeney, et al. 2015a; Keeney, et al. 2014). This is supported by homology to genes with known functions in cell cycle dynamics (Popesco, et al. 2006; Thornton and Woods 2009), relevant spatial and temporal expression patterns (Keeney, et al. 2015a) and an effect on the proliferation of neuroblastoma cell cultures (Vandepoele, et al. 2008). However, a direct effect of variation in DUF1220 domain number on neural proliferation has not been demonstrated (Keeney, et al. 2015b).

If DUF1220 domains do regulate neurogenesis, they would be expected to co-evolve with prenatal brain growth, as cortical neurogenesis is restricted to prenatal development (Bhardwaj, et al. 2006). Our results instead suggest a robust and specific relationship with postnatal brain development. Existing data on DUF1220 domain function suggest two potential roles that may explain this association: i) a contribution to axonogenesis via initiating and stabilizing microtubule growth in dendrites; and ii) a potential role in apoptosis during brain maturation. Both hypotheses are consistent with the reported association between variation in DUF1220 dosage and ASD (Davis, et al. 2014). Indeed, an emphasis on postnatal brain growth is potentially more relevant for ASD which develops postnatally, accompanied by a period of accelerated brain growth in early postnatal development (Courchesne, et al. 2001).

Microtubule assembly is essential for dendritic growth and axonogenesis (Conde and Cáceres 2009). PDE4DIP, which contains the ancestral DUF1220 domain, has known functions in microtubule nucleation, growth, and cell migration (Roubin, et al. 2013). There is also evidence NBPF1 interacts with a key regulator of Wnt signalling (Vandepoele, et al. 2010), which has important roles in neuronal differentiation, dendritic growth and plasticity (Inestrosa and Varela-Nallar 2015). Consistent with this function, DUF1220 domains are highly expressed in the cell bodies and dendrites of adult neurons (Popesco, et al. 2006). A role for DUF1220 domains in synaptogenesis could potentially explain the association with ASD severity (Davis, et al. 2014). ASD is associated with abnormalities in cortical minicolumns (Casanova, et al. 2002) and cortical white matter (Courchesne, et al. 2001; Hazlett, et al. 2005), both of which suggest a disruption of normal neuronal maturation (Courchesne and Pierce 2005; Minshew and Williams 2007).

Alternatively, *NBPF* genes are also known to interact with NF-κB (Zhou, et al. 2013), a transcription factor implicated in tumor progression, with a range of roles including apoptosis and inflammation (Karin and Lin 2002; Perkins 2012). Postnatal apoptosis has a significant influence on brain growth (Kuan, et al. 2000; Madden, et al. 2007; Polster, et al. 2003), including regulating neuronal density (Sanno, et al. 2010), and apoptotic genes may have been targeted by selection in relation to primate brain expansion (Vallender and Lahn 2006). Disruption of apoptosis causes microcephaly (Poulton, et al. 2011) potentially explaining the association between DUF1220 dosage and head circumference (Dumas, et al. 2012). The association of NF-κB with inflammatory diseases (Tak and Firestein 2001) is also intriguing, given the growing evidence that the inflammatory response is linked to the risk and severity of ASD (Depino 2013; Meyer, et al. 2011).

Finally, these results add further evidence that many of the genetic changes that contribute to human evolution will be based on the continuation or exaggeration of conserved gene-phenotype associations that contribute to primate brain evolution more broadly (Montgomery, et al. 2011; Scally, et al. 2012). Understanding the commonalities between human and non-human primate brain evolution is therefore essential to understand the genetic differences that contribute the derived aspects of human evolution.

Chapter 7

Discussion

In this thesis, I have characterised evolutionary changes in genomic architecture via gene duplication, loss and translocation. I sought to test whether dosage-sensitive genes are preferentially dosage-compensated on the avian sex chromosomes. In chapter 2, I showed that dosage-compensation on the avian Z chromosome specifically affects dosage-sensitive genes, suggesting that gene dose is an important factor in the evolution of sex chromosomes. I asked if it is possible to use *de novo* RNA-Seq assemblies for the detection of lineage-specific paralogs. In chapter 3, I developed a bioinformatics pipeline to reconstruct gene family history using *de novo* RNA-Seq assemblies and described pitfalls and shortcomings of doing so with currently available methods. Additionally, I sought to test whether the genomic distribution of mito-nuclear genes is shaped by different selective pressures in males and females. In chapters 4 and 5, I investigated the distribution of mito-nuclear genes in multiple species with different sex chromosome systems and only detected a paucity of mito-nuclear genes on the mammalian X chromosome and in *C. elegans*. In all other studied sex chromosome systems, I did not find a significant underrepresentation. My synteny analysis of the human X chromosome also indicated that the paucity of mito-nuclear genes on the mammalian X chromosome likely predates the evolution of the mammalian sex chromosome system. Finally, in chapter 6, I investigated a case study of rapid protein domain duplication to test whether increases in protein domain dose correspond to phenotypic evolution across species.

First, I will discuss the results presented in chapters 2-6 along with some specific issues and limitations of the analyses. Subsequently, I will discuss the limitations and biases of different types of biological data used in this thesis, with the aim of highlighting potential pitfalls and avenues for improvement. Specifically, I will discuss current challenges in the analysis of RNA-Seq data, such as how differences in annotation quality can bias analyses. Finally, I will briefly discuss the importance of scientific computing for modern biology, best practices for bioinformatics analyses and future directions of the field.

Dosage compensation and whole genome duplications

My first aim in this thesis was to understand if the dosage compensation mechanism balances the gene expression of dosage-sensitive genes on the avian sex chromosomes. In chapter 2, I used ohnologs, genes duplicated in a whole genome duplication, as proxies for dosage-sensitive genes and showed that they are indeed preferentially dosage-compensated. This is important because it shows that the dosage compensation mechanism in birds is not 'incomplete' in the sense that it does not balance out dosage effects. Rather, it effectively balances the expression on a gene-by-gene basis. These results are similar to the XY sex chromosome system in mammals, where dosage-sensitive genes on the X chromosome are also dosage-balanced (Pessia, et al. 2012). Taken together, these results suggest that dosage compensation mechanisms evolved in order to balance out the expression of a small subset of dosage-sensitive genes located on the sex chromosomes.

Previous investigations in several different bird species confirmed the lack of a global dosage compensation mechanism (Naurin, et al. 2011; Uebbing, et al. 2015; Uebbing, et al. 2013; Wang, et al. 2014; Wright, et al. 2015b), and my results are consistent with these observations. In addition to studies in single species, several phylogenetic analyses of the evolution of dosage compensation in birds have been conducted in recent years (Julien, et al. 2012; Wang, et al. 2014; Wright, et al. 2015b). These studies confirm the absence of a global dosage compensation mechanism in birds and also showed that the hypermethylated region on the male Z chromosomes is not an area of nascent dosage compensation (Wright, et al. 2015b). Phylogenetic analyses are especially important because they enable the reconstruction of ancestral expression levels. For example, the reduced Z chromosome expression in chicken was interpreted as a potential sign of Z chromosome inactivation (Livernois, et al. 2013). I also recovered a lower level of Z expression compared to the autosomes but my analysis of allele-specific expression failed to recover evidence for the partial inactivation of one Z chromosome. Rather than being caused by partial inactivation, the lower Z expression could therefore be a result of lower ancestral expression level (Julien, et al. 2012).

When interpreting and generalising the results presented in chapter 2, some methodological factors need to be considered. It is important to recognise that the detection of ohnologs in vertebrate genomes remains challenging. All tools for the detection of ohnologs depend on the analysis of preserved gene order (synteny) among paralogs to distinguish single-gene duplicates from WGD. Large intra-genomic rearrangements may complicate these analyses, and result in the underestimation of the number of ohnologs. Avian genomes, however, are relatively stable and compact, with fewer repeats and more coding DNA, compared to other amniotes (Ellegren 2005; Hillier, et al. 2004; Organ, et al. 2007), making these issues less problematic. The detection of ohnologs may also depend on the selection of one or more outgroup species, to distinguish between ohnologs and genes that were duplicated before the WGD events. The specific outgroup selected can influence the number of ohnologs (Makino and McLysaght 2010) and also relies on the *a priori* assumption that a WGD did not take place in this lineage. The pipeline used in the OhnologsDB tries to mitigate this issue by using multiple outgroups. It also provides sets of ohnologs with varying strictness, and my results are consistent with both the ‘relaxed’ and ‘strict’ set.

Inferring gene family evolution from RNA-Seq data

Comparative analyses of gene expression are important because the divergence of gene expression following duplication may be a key aspect for the maintenance and evolution of duplicates and likely plays a major role in adaptation. I sought to identify paralogs in six different bird species using primarily RNA-Seq data. My aim was to explore whether these data can be used in conjunction with tools that usually rely on genomic data and subsequently investigate the gene expression divergence of the detected paralogs. In chapter 3, I developed a pipeline that used *de novo* RNA-Seq assembly data in combination with DNA-Seq data across a range of bird species.

Until recently, comparative studies of gene expression have relied on the availability of reference genomes (e.g. Brawand, et al. 2011). This limits the analyses to species with available genomic resources; however, many species studied to date are not yet sequenced. Other comparative studies of gene expression have used RNA-Seq based assemblies and subsequently aligned RNA-Seq based transcripts to a well-annotated reference (Harrison, et al. 2015; Wright, et al. 2015a; Yang and Smith 2013). By doing so, it is impossible to detect any genes that are not present in the reference and thus any form of gene duplications in other lineages. This approach is therefore not possible across large or even moderate phylogenetic distances. In contrast, the bioinformatics pipeline presented in chapter 3 does not primarily rely on genomic data. Previous studies found that bird genomes are relatively stable compared to mammals (Hillier, et al. 2004; Jarvis, et al. 2014). With a low duplication rate of ca. 0.2 single gene duplications per million years and a sparsity of young paralogs (Toups, et al. 2011), the duplication rate estimated using RNA-Seq data is roughly two to four times higher. This indicates that even a small and strictly filtered dataset obtained using this approach is likely to contain a higher number of false positives. There are several methodological reasons that potentially contribute to the inability to reliably detect paralogs using RNA-Seq data.

Regardless of the software tool used, *de novo* RNA-Seq assemblies generate a large number of transcripts, many of which have very low expression levels and probably little or no biological significance (Raj and van Oudenaarden 2008). In order to use this kind of dataset for the reconstruction of gene family histories, I reduced the number of transcripts by using strict filtering steps, including gene expression thresholds, family structure and gene tree analyses. I also used a threshold of > 2 RPKM that was previously used in analyses of this dataset (Harrison, et al. 2015; Wright, et al. 2015a). After filtering, the number of sequences with a valid open reading frame is comparable to genomic data. Despite the gene expression filter, the quality and completeness of gene models created from RNA-Seq data is much lower in comparison to DNA-Seq based data. This placed major limitations of the ability to reconstruct gene families because it resulted in the generation of large amounts of losses and singletons. High numbers of losses are indicative of fragmented assemblies and singletons likely represent low quality data with questionable biological relevance.

I also employed a family-structure filter to detect lineage-specific paralogs. However, alignments between many paralog pairs showed very low sequence identity. Gene tree analyses subsequently revealed that most paralogs do not form a monophyletic group compared to an outgroup. This could indicate a very rapid change in coding sequence, but is more likely the result of misidentification of paralogs due to the high amount of noise in the data. For duck and turkey, both RNA-Seq and DNA-Seq data was available and lineage-specific paralogs identified in these species could be validated using the Ensembl database (Cunningham, et al. 2015). All pairs of paralogs in these groups are also identified in the Ensembl database; however, none of them are marked as lineage-specific and all Ensembl estimated duplication dates preceded the evolution of the avian clade. This is likely the result of missing data in many RNA-Seq assemblies and further evidence that the low quality of the RNA-Seq gene models limits the applicability of this approach.

In the future, sequencing technologies that generate reads with the length of several kilobases, such as nanopore-based sequencing, could be used to generate draft reference genomes very efficiently (Feng, et al. 2015). Longer reads reduce the assembly time significantly because of the reduced size of the De-Bruijn graph. This would make it attractive to sequence a draft genome first, before performing any gene expression analyses and could possibly render *de novo* transcriptome assemblies obsolete. However, estimating gene expression levels directly using long read technology is currently not possible due to a very limited read depth. In combination, long-read sequencing techniques and classic RNA-Seq short-read sequencing could enable analyses of paralog expression divergence over a larger phylogenetic range.

Gene movement and mitochondrial interactions

The unequal inheritance of the mitochondrial genome in comparison to the autosomes leads to contrasting predictions regarding the location of mito-nuclear genes in males and females. In females, an overrepresentation of mito-nuclear genes on the X chromosome could be a favourable genomic location because X-linked genes are co-transmitted with the mitochondrial genome two thirds of the time (Brandvain and Wade 2009). An overrepresentation of X-linked mito-nuclear genes would provide more scope for effective female-specific selection. The opposite is true for males: X-linked mito-nuclear genes spend even less time in males, which would reduce the opportunity for male-specific selection to favour mutations that counteract detrimental male-specific effects. This could drive mito-nuclear genes off the X chromosome (Drown, et al. 2012; Rice 1984; Werren 2011). In line with this prediction, an underrepresentation of mito-nuclear genes on the sex chromosomes has been observed on the mammalian X chromosome (Drown, et al. 2012). However, the underrepresentation of mito-nuclear genes was only analysed in one sex chromosome system that is largely conserved across multiple species. In chapter 4, we investigated the genomic distribution of mito-nuclear genes in multiple species with different sex chromosome systems. We confirmed a paucity of mito-nuclear genes in therian mammals and found a similar pattern in *C. elegans*. However, all other species studied did not show a significant under- or overrepresentation of mito-nuclear genes on the sex chromosomes. This indicates that neither hypotheses based on sexual conflict nor sexual selection in female heterogametic systems can fully explain the distribution of mito-nuclear genes across multiple, independent sex chromosome systems.

Consequently, we developed an alternative hypothesis to explain the paucity of mito-nuclear genes on the therian X chromosome. Mito-nuclear interactions precede the formation of most sex chromosome systems, and it is possible that the observed underrepresentation of mito-nuclear genes is the result of an ancestral under representation of the autosomes, which subsequently evolved into sex chromosomes. I tested this hypothesis by reconstructing the syntenic relationship of the human X chromosome to the ancestral chromosomes in *G. gallus* (chicken) and *O. anatinus* (platypus) and found that the paucity of mito-nuclear genes was already present in the ancestral regions. Additionally, there was no significant movement off the X chromosome, suggesting that the underrepresentation is an ancestral trait, not the result of sexual conflict after the origin of the sex chromosomes. Given these results, it is likely that the observed underrepresentation of mito-nuclear genes on the mammalian X chromosomes (Drown, et al. 2012) is just a chance event and not caused by selection for mito-nuclear gene movement off the X chromosome in males.

DUF1220 domain dose increase

The amplification of protein domains could have played an important role in human evolution (Emerson and Thomas 2009; Popesco, et al. 2006). However, robust tests associating protein domain gain with phenotypic data in a phylogenetic framework are still lacking. We thus sought to test if there is an association between an increase in protein domain copy number and phenotypic change. In chapter 5, we investigated the association between domain dose and brain evolution in primates. In order to analyse the DUF1220 protein domain association with phenotypic data across a hierarchical phylogenetic tree, I developed a pipeline for the identification of DUF1220 domains. I used a DUF1220 Hidden Markov Model (HMM) obtained from the PFAM database to detect domains across species of primate which last shared a common ancestor ca. 76 million years ago (Hedges, et al. 2015), then incorporated all detected sequences and built a specific nucleotide HMM. Previous analyses were based on BLAST/BLAT, where a single sequence from one species is used to detect homologous sequences across a range of species. HMMs built from a range of different species have the advantage of increased accuracy of detection and reduce the phylogenetic bias that is inherent in using one single sequence (Terrapon, et al. 2012). My results indicate that custom-built HMMs detect protein domains more accurately and that the phylogenetic bias is significantly reduced compared to BLAST/BLAT-based methods.

The analyses of the DUF1220 domains confirm a strong association with brain mass, and in particular neocortex volume (Dumas, et al. 2012; Keeney, et al. 2014). We also demonstrated that DUF1220 domains are associated with postnatal and not prenatal brain growth, which contrasts with previous work that hypothesised a specific association with prenatal neurogenesis (Dumas and Sikela 2009; Keeney, et al. 2015a; Keeney, et al. 2014) based on homology (Popesco, et al. 2006) and temporal expression data (Keeney, et al. 2015a). Instead, the analyses are consistent with a potential role for DUF1220 containing genes in microtubule growth (Roubin, et al. 2013) or apoptosis through direct interactions with NF-κB (Zhou, et al. 2013), processes critical for postnatal brain development. This is important, because the results fit with a proposed association of DUF1220 and Autism Spectrum Disorder (ASD) (Davis, et al. 2014). ASD is thought to develop postnatally (Courchesne, et al. 2001) rather than prenatally, and an increasing body of evidence suggests a role of inflammatory responses (Depino 2013; Meyer, et al. 2011).

If DUF1220 domain number does contribute to the evolution of postnatal brain growth, this contrasts with results of previously studied candidate genes with known roles in neurogenesis that co-evolve with prenatal brain growth (Montgomery, et al. 2011). This suggests a two-component model of brain evolution where selection targets one set of genes to bring about an increase in neuron number (Montgomery, et al. 2011; Montgomery and Mundy 2012a; Montgomery and Mundy 2012b), and an independent set of genes to optimise neurite growth and connectivity (Charrier, et al. 2012). DUF1220 domain containing Neuro Blastoma Break Point Family (*NBPF)* genes may fall into the latter category. This two-component model is consistent with comparative analyses indicating that pre- and postnatal brain development evolve independently, and must therefore be relatively free of reciprocal pleiotropic effects (Barton and Capellini 2011).

These analyses of DUF1220 across primates also revealed that there is a strong annotation bias across primate genomes. In comparison to the comparatively well-annotated human genome, other genomes are only available as scaffolds with low-quality gene model predictions that are rarely updated. This annotation bias can lead to the overestimation of human lineage-specific trends because higher quality assemblies provide a more accurate picture of the genomic architecture in regions with large number of protein domain repeats. In order to account for the differences in annotation quality, I analysed the DUF1220 counts across a range of different types of data (proteins, nucleotides, exonic), which revealed the robustness of the observed pattern. Additionally, I excluded DUF1220 present in pseudogenes because pseudogenes are not translated, and counting protein domains present in pseudogenes may therefore not be biologically meaningful. Accounting for these biases is crucial when examining trends within and between species, as these differences can, for example, bias studies of gene expression (Zhao and Zhang 2015).

Finally, it is important to account for the phylogenetic non-independence when analysing the association of count data with phenotypic traits (Carvalho, et al. 2006; Felsenstein 1985). Failure to do so can result in ‘phylogenetic pseudo replication’ (Garland 2001). We used phylogenetic generalised linear mixed models and a Bayesian approach implemented in MCMCglmm (Hadfield 2010) to investigate gene-phenotype interactions while correcting for phylogenetic relatedness. Previous studies analysing DUF1220 protein domains ignored this issue (Dumas, et al. 2012; Keeney, et al. 2015b); doing so can result in an inflation of statistical significance (Carvalho, et al. 2006). This is also an issue when comparing gene expression values between species (Dunn, et al. 2013) and could bias some of the results of recent comparative studies (e.g.Brawand, et al. 2011) or when comparisons are made using homologous structures across a range of closely related species. In the future, phylogenetic analyses should become more prevalent as comparisons between small numbers of species make the interpretation of the results challenging.

Analysing comparative transcriptome data: general lessons

The availability of transcriptomic data across a range of organisms has facilitated many analyses of gene expression. The design of RNA-Seq experiments necessitates using many different combinations of tools, depending on the focus of the study (Conesa, et al. 2016). Here, I would like to highlight three issues that I consider important when it comes to the calculation of differential expression using RNA-Seq.

First, RNA-Seq experiments still underutilise biological replicates in favour of higher sequencing depths, despite a clear increase in statistical power when using multiple samples (Liu, et al. 2014; Robles, et al. 2012). All RNA-Seq data analysed in chapters 2 and 3 use four to six biological replicates for every tissue in order to resolve this issue, except for turkey where only two female spleen samples could be obtained due to sampling constraints. Four replicates are sufficient for the detection of differentially expressed genes with a power of >80%, a sequencing depths of 5 million reads and an FDR level of <0.05 (Liu, et al. 2014). In chapter 2, on average 17 million mappable reads were available per sample, which results in enough power to reliably identify differential gene expression. However, more biological replicates will increase the power even further and should always be preferred over higher sequencing depth (Liu, et al. 2014).

Secondly, RNA-Seq data needs to be normalised when comparing expression data between conditions and across libraries. Normalisation methods do not only account for differences in library size and gene length (CPM or RPKM), but also for other technical variables. Recently, comparisons on the effectiveness of different normalisation methods have become available (Lin, et al. 2016). I used the TMM method implemented in edgeR (Robinson, et al. 2010), which is one of the two recommended normalisation methods (Lin, et al. 2016). TMM is sensitive to filtering of genes and I only filtered out lowly expressed genes after normalisation, as is recommended by Lin, et al. (2016). For comparisons of the same gene across conditions I used Counts Per Million (CPM) values and Reads Per Kilobase of transcript per Million mapped reads (RPKM) values when it was necessary to correct for gene length. One additional limitation of the normalisation methods used is that they assume the majority of genes are not differentially expressed; however, this assumption is usually violated in gonadal tissue, where the distribution of gene expression can be fundamentally different. In this case, normalisation may mistake biological variation with technical variation and equalise two different distribution of gene expression. I used multiple tissues that included both somatic and gonadal tissue in order to account for this issue.

Finally, differences in tissue scaling, in particular allometric scaling, between conditions can significantly bias the inference of differential gene expression (Harrison, et al. 2015). This issue is still not widely recognised but Montgomery and Mank (2016) suggest that using fold-change cut-offs in conjunction with other statistical tests for differential expression can mitigate some of the problems. These issues are of central importance as they affect the comparability of datasets. Any comparative analysis based on non-comparable data will produce biased results. Improved methods of sampling and data processing will be central to the future use of RNA-Seq in evolutionary studies.

Annotation and assembly quality differences

Comparisons between sequence data are commonly used to infer evolutionary relationships between genes. These comparisons depend on the quality of genome or transcriptome assemblies and annotations across the range of study species. I encountered issues with the gene annotation and assembly quality in several of my chapters. In chapter 5, for example, the syntenic regions between the human X chromosome and the platypus genome are scattered across multiple UltraContigs (**Figure 5.1**). These UltraContigs do not constitute ‘real’ chromosomes and are a result of an incomplete assembly. In comparison to the human genome, the assembly quality of the platypus genome is much lower, which could bias the analysis. In order to mitigate this issue, I used a second comparison to the well-annotated chicken genome. In this case I recovered similar results, but in many analyses genome quality may have a major influence on the patterns detected in evolutionary analyses.

Further effects of the issue of gene annotation quality can be found in chapter 6. In this chapter, I compared the DUF1220 number in human peptides with counts for chimp and rhesus macaque across different Ensembl database versions. This shows that differences in annotation quality can impact comparative results and that this effect may lead to significant differences, not because of evolutionary changes but solely because of annotation bias. Differences in the quality of gene models between RNA-Seq and DNA-Seq based data also affected the analyses presented in chapter 3. Gene models from RNA-Seq based *de novo* assemblies often result in protein sequences that lack any similar sequences in closely related species; this can negatively affect the reconstruction of gene families.

Finally, Gene Ontology terms (Ashburner, et al. 2000) for comparisons of functional differences between species should be conducted very carefully, as they suffer from various annotation biases (Altenhoff, et al. 2012; Schnoes, et al. 2013). These biases make GO terms potentially unsuitable to test specific hypotheses, such as the ortholog conjecture (Chen and Zhang 2012).

General limitations in bioinformatics analyses

The increasing amount of biological data requires many biologists to use computational tools and methods to analyse their data. The computational analysis of biological data is changing rapidly and requires a set of non-biology related skills, most importantly programming or scripting. There is a clear need for best practices in scientific computing (Wilson, et al. 2014), software development and scripting (Leprevost, et al. 2014), improved training in computational skills and the way these steps are reported in published papers. In this section, I will briefly discuss these challenges and advances made in the field today.

The large number of computational analyses used to tackle biological questions does not fit well into the methods section of a traditionally formatted paper. An accurate description of the methods used should allow other scientists to fully understand all steps taken, and ultimately allow them to reproduce the study. For this reason, all database versions used, versions of software and all defined parameters should be described, but often they are not. In addition, many studies rely on custom-written software, which should be freely available in a public, version controlled repository, such as GitHub (Wilson, et al. 2014). This should be a requirement for publication because programming code is as central to the analyses as the data, open access to which is now routinely demanded by journals. However, for many published studies source code is not made available, thereby increasing the difficulty of reproducing studies. Any repository available on GitHub that contains scientifically relevant source code can be assigned a citable Digital Object Identifier (DOI), which could provide a better incentive to make the source code available. A study by Kidwell, et al. (2016) suggested that a simple badge system can also help to encourage more transparent and open practices. Recent changes, for example in the Data Policy of the Public Library of Science (PLoS) journals, also require the open accessibility of all collected data and PloS encourages researchers to make source code available under an open source license.

Even if all data and software are made available, it can be hard to reproduce results that rely on the correct execution of different software packages and custom written scripts. Recent software, such as Docker and LXC, allows the creation of container images that include all software needed to rerun a computational pipeline with relative independence from the host system. Future studies could provide a single (or multiple) Docker container images, which allow the exact replication of computational workflows. In combination with defined ontologies, such as the Common Workflow Language (http://www.commonwl.org/) computational workflows can be created, shared and reproduced. In summary, to produce the best science we have to continue establishing systems that ensure transparency, reproducibility and reward the use of best practices in scientific computing and bioinformatics.

References

Unavailable fields are indicated as '–'

Abascal F, Zardoya R, Telford MJ (2010). TranslatorX: Multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Research* 38(suppl 2): W7-W13.
doi: 10.1093/nar/gkq291

Albertin W, Marullo P (2012). Polyploidy in fungi: Evolution after whole-genome duplication. *Proceedings of the Royal Society of London B: Biological Sciences* 279(1738): 2497-2507.
doi: 10.1098/rspb.2012.0434

Albritton SE, et al. (2014). Sex-biased gene expression and evolution of the X chromosome in nematodes. *Genetics* 197(3): 865-883.
doi: 10.1534/genetics.114.163311

Altenhoff AM, Gil M, Gonnet GH, Dessimoz C (2013). Inferring Hierarchical Orthologous Groups from orthologous gene pairs. *PLoS ONE* 8(1): e53786.
doi: 10.1371/journal.pone.0053786

Altenhoff AM, et al. (2015). The OMA orthology database in 2015: Function predictions, better plant support, synteny view and other improvements. *Nucleic Acids Research* 43(D1): D240-D249.
doi: 10.1093/nar/gku1158

Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C (2012). Resolving the ortholog conjecture: Orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS Computational Biology* 8(5): e1002514.
doi: 10.1371/journal.pcbi.1002514

Altschul SF, et al. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology* 215(3): 403-410.
doi: 10.1016/S0022-2836(05)80360-2

Anders S, Pyl PT, Huber W (2015). HTSeq - a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31(2): 166-169.
doi: 10.1093/bioinformatics/btu638

Andersson DI, Hughes D (2009). Gene amplification and adaptive evolution in bacteria. *Annual Review of Genetics* 43(1): 167-195.
doi: 10.1146/annurev-genet-102108-134805

Andolfatto P (2001). Contrasting patterns of X-linked and autosomal nucleotide variation in *Drosophila melanogaster* and *Drosophila simulans*. *Molecular Biology and Evolution* 18(3): 279-290.
doi: -

Aparicio S, et al. (2002). Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* 297(5585): 1301-1310.
doi: 10.1126/science.1072104

Arnold C, Matthews LJ, Nunn CL (2010). The 10kTrees website: A new online resource for primate phylogeny. *Evolutionary Anthropology: Issues, News, and Reviews* 19(3): 114-118.
doi: 10.1002/evan.20251

Arunkumar KP, Mita K, Nagaraju J (2009). The silkworm Z chromosome is enriched in testis-specific genes. *Genetics* 182(2): 493-501.
doi: 10.1534/genetics.108.099994

Ashburner M, et al. (2000). Gene Ontology: Tool for the unification of biology. *Nature Genetics* 25(1): 25-29.
doi: 10.1038/75556

Assis R, Bachtrog D (2013). Neofunctionalization of young duplicate genes in *Drosophila*. *Proceedings of the National Academy of Sciences of the United States of America* 110(43): 17409-17414.
doi: 10.1073/pnas.1313759110

Aury J-M, et al. (2006). Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature* 444(7116): 171-178.
doi: 10.1038/nature05230

Bachtrog D, et al. (2011). Are all sex chromosomes created equal? *Trends in Genetics* 27(9): 350-357.
doi: 10.1016/j.tig.2011.05.005

Bachtrog D, et al. (2014). Sex determination: Why so many ways of doing it? *PLoS Biology* 12(7): e1001899.
doi: 10.1371/journal.pbio.1001899

Baker BS, Belote JM (1983). Sex determination and dosage compensation in *Drosophila melanogaster*. *Annual Review of Genetics* 17(1): 345-393.
doi: 10.1146/annurev.ge.17.120183.002021

Band MR, et al. (2000). An ordered comparative map of the cattle and human genomes. *Genome Research* 10(9): 1359-1368.
doi: 10.1101/gr.145900

Barton RA, Capellini I (2011). Maternal investment, life histories, and the costs of brain growth in mammals. *Proceedings of the National Academy of Sciences of the United States of America* 108(15): 6169-6174.
doi: 10.1073/pnas.1019140108

Barton RA, Harvey PH (2000). Mosaic evolution of brain structure in mammals. *Nature* 405(6790): 1055-1058.
doi: 10.1038/35016580

Barton RA, Venditti C (2014). Rapid evolution of the cerebellum in humans and other great apes. *Current Biology* 24(20): 2440-2444.
doi: 10.1016/j.cub.2014.08.056

Betrán E, Thornton K, Long M (2002). Retroposed new genes out of the X in *Drosophila*. *Genome Research* 12(12): 1854-1859.
doi: 10.1101/gr.604902

Beukeboom LW, Perrin N. (2014). The evolution of sex determination: Oxford University Press, USA.

Bhardwaj RD, et al. (2006). Neocortical neurogenesis in humans is restricted to development. *Proceedings of the National Academy of Sciences of the United States of America* 103(33): 12564-12568.
doi: 10.1073/pnas.0605177103

Birchler JA, Bhadra U, Bhadra MP, Auger DL (2001). Dosage-dependent gene regulation in multicellular eukaryotes: Implications for dosage compensation, aneuploid syndromes, and quantitative traits. *Developmental Biology* 234(2): 275-288.
doi: 10.1006/dbio.2001.0262

Birchler JA, Riddle NC, Auger DL, Veitia RA (2005). Dosage balance in gene regulation: biological implications. *Trends in Genetics* 21(4): 219-226.
doi: 10.1016/j.tig.2005.02.010

Birchler JA, Veitia RA (2012). Gene balance hypothesis: Connecting issues of dosage sensitivity across biological disciplines. *Proceedings of the National Academy of Sciences of the United States of America* 109(37): 14746-14753.
doi: 10.1073/pnas.1207726109

Birol I, et al. (2009). *De novo* transcriptome assembly with ABySS. *Bioinformatics* 25(21): 2872-2877.
doi: 10.1093/bioinformatics/btp367

Björklund Å, Ekman D, Elofsson A (2006). Expansion of protein domain repeats. *PLoS Computational Biology* 2(8): e114.
doi: 10.1371/journal.pcbi.0020114

Björklund ÅK, et al. (2005). Domain rearrangements in protein evolution. *Journal of Molecular Biology* 353(4): 911-923.
doi: 10.1016/j.jmb.2005.08.067

Björklund ÅK, Light S, Sagit R, Elofsson A (2010). Nebulin: A study of protein repeat evolution. *Journal of Molecular Biology* 402(1): 38-51.
doi: 10.1016/j.jmb.2010.07.011

Blakeslee AF, Belling J, Farnham ME (1920). Chromosomal duplication and mendelian phenomena in *Datura* mutants. *Science* 52(1347): 388-390.
doi: 10.1126/science.52.1347.388

Blanc G, Wolfe KH (2004). Functional divergence of duplicated genes formed by polyploidy during arabidopsis evolution. *The Plant Cell* 16(7): 1679-1691.
doi: 10.1105/tpc.021410

Blomme T, et al. (2006). The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biology* 7(5): 1-12.
doi: 10.1186/gb-2006-7-5-r43

Bolger AM, Lohse M, Usadel B (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15): 2114-2120.
doi: 10.1093/bioinformatics/btu170

Bond J, et al. (2005). A centrosomal mechanism involving *CDK5RAP2* and *CENPJ* controls brain size. *Nature Genetics* 37(4): 353-355.
doi: 10.1038/ng1539

Bornberg-Bauer E, Albà MM (2013). Dynamics and adaptive benefits of modular protein evolution. *Current Opinion in Structural Biology* 23(3): 459-466.
doi: 10.1016/j.sbi.2013.02.012

Bornberg-Bauer E, et al. (2005). The evolution of domain arrangements in proteins and interaction networks. *Cellular and Molecular Life Sciences CMLS* 62(4): 435-445.
doi: 10.1007/s00018-004-4416-1

Brandvain Y, Wade MJ (2009). The functional transfer of genes from the mitochondria to the nucleus: The effects of selection, mutation, population size and rate of self-fertilization. *Genetics* 182(4): 1129-1139.
doi: 10.1534/genetics.108.100024

Braun RE, et al. (1989). Genetically haploid spermatids are phenotypically diploid. *Nature* 337(6205): 373-376.
doi: 10.1038/337373a0

Brawand D, et al. (2011). The evolution of gene expression levels in mammalian organs. *Nature* 478(7369): 343-348.
doi: 10.1038/nature10532

Bridges CB (1936). The Bar "gene" a duplication. *Science* 83(2148): 210-211.
doi: 10.1126/science.83.2148.210

Bridges CB (1914). Direct proof through non-disjunction that the sex-linked genes of *Drosophila* are borne by the X-chromosome. *Science* 40(1020): 107-109.
doi: 10.1126/science.40.1020.107

Brunet FG, et al. (2006). Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Molecular Biology and Evolution* 23(9): 1808-1816.
doi: 10.1093/molbev/msl049

Buchman JJ, et al. (2010). Cdk5rap2 interacts with pericentrin to maintain the neural progenitor pool in the developing neocortex. *Neuron* 66(3): 386-402.
doi: 10.1016/j.neuron.2010.03.036

Buljan M, Bateman A (2009). The evolution of protein domain families. *Biochemical Society Transactions* 37(4): 751-755.
doi: 10.1042/BST0370751

Bull JJ. (1983). Evolution of sex determining mechanisms: The Benjamin/Cummings Publishing Company, Inc.

Burt DW, et al. (1999). The dynamics of chromosome evolution in birds and mammals. *Nature* 402(6760): 411-413.
doi: 10.1038/46555

Byrne KP, Wolfe KH (2007). Consistent patterns of rate asymmetry and gene loss indicate widespread neofunctionalization of yeast genes after whole-genome duplication. *Genetics* 175(3): 1341-1350.
doi: 10.1534/genetics.106.066951

Camus MF, Clancy DJ, Dowling DK (2012). Mitochondria, maternal inheritance, and male aging. *Current Biology* 22(18): 1717-1721.
doi: 10.1016/j.cub.2012.07.018

Capellini I, Venditti C, Barton Robert A (2011). Placentation and maternal investment in mammals. *The American Naturalist* 177(1): 86-98.
doi: 10.1086/657435

Carvalho P, Diniz-Filho JAF, Bini LM (2006). Factors influencing changes in trait correlations across species after using phylogenetic independent contrasts. *Evolutionary Ecology* 20(6): 591-602.
doi: 10.1007/s10682-006-9119-7

Casanova MF, et al. (2002). Minicolumnar pathology in dyslexia. *Annals of Neurology* 52(1): 108-110.
doi: -

Charlesworth B (1996). The evolution of chromosomal sex determination and dosage compensation. *Current Biology* 6(2): 149-162.
doi: 10.1016/S0960-9822(02)00448-7

Charlesworth B (1991). The evolution of sex chromosomes. *Science* 251(4997): 1030-1033.
doi: 10.1126/science.1998119

Charlesworth B (1978). Model for evolution of Y chromosomes and dosage compensation. *Proceedings of the National Academy of Sciences of the United States of America* 75(11): 5618-5622.
doi: -

Charlesworth B (1998). Sex chromosomes: Evolving dosage compensation. *Current Biology* 8(25): R931-R933.
doi: 10.1016/S0960-9822(98)00013-X

Charlesworth B, Coyne JA, Barton NH (1987). The relative rates of evolution of sex chromosomes and autosomes. *American Naturalist* 130(1): 113-146.
doi: -

Charlesworth D (2013). Plant sex chromosome evolution. *Journal of Experimental Botany* 64(2): 405-420.
doi: 10.1093/jxb/ers322

Charlesworth D, Charlesworth B, Marais G (2005). Steps in the evolution of heteromorphic sex chromosomes. *Heredity* 95(2): 118-128.
doi: 10.1038/sj.hdy.6800697

Charrier C, et al. (2012). Inhibition of *SRGAP2* function by its human-specific paralogs induces neoteny during spine maturation. *Cell* 149(4): 923-935.
doi: 10.1016/j.cell.2012.03.034

Chen S, et al. (2014). Whole-genome sequence of a flatfish provides insights into ZW sex chromosome evolution and adaptation to a benthic lifestyle. *Nature Genetics* 46(3): 253-260.
doi: 10.1038/ng.2890

Chen X, Zhang J (2012). The ortholog conjecture is untestable by the current gene ontology but is supported by RNA sequencing data. *PLoS Computational Biology* 8(11): e1002784.
doi: 10.1371/journal.pcbi.1002784

Chothia C, Gough J, Vogel C, Teichmann SA (2003). Evolution of the protein repertoire. *Science* 300(5626): 1701-1703.
doi: 10.1126/science.1085371

Clark AG (1985). Natural selection with nuclear and cytoplasmic transmission. II. Tests with *Drosophila* from diverse populations. *Genetics* 111(1): 97-112.
doi: -

Clark AG, Lyckegaard EM (1988). Natural selection with nuclear and cytoplasmic transmission. III. Joint analysis of segregation and mtDNA in *Drosophila melanogaster*. *Genetics* 118(3): 471-481.
doi: -

Cock PJA, et al. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25(11): 1422-1423.
doi: 10.1093/bioinformatics/btp163

Coghlan A, et al. (2005). Chromosome evolution in eukaryotes: A multi-kingdom perspective. *Trends in Genetics* 21(12): 673-682.
doi: 10.1016/j.tig.2005.09.009

Conde C, Cáceres A (2009). Microtubule assembly, organization and dynamics in axons and dendrites. *Nature Reviews Neuroscience* 10(5): 319-332.
doi: 10.1038/nrn2631

Conesa A, et al. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology* 17(1): 1-19.
doi: 10.1186/s13059-016-0881-8

Connallon T (2007). Adaptive protein evolution of X-linked and autosomal genes in *Drosophila*: implications for faster-X hypotheses. *Molecular Biology and Evolution* 24(11): 2566-2572.
doi: 10.1093/molbev/msm199

Conrad T, Akhtar A (2012). Dosage compensation in *Drosophila melanogaster*: Epigenetic fine-tuning of chromosome-wide transcription. *Nature Reviews Genetics* 13(2): 123-134.
doi: 10.1038/nrg3124

Coolon JD, et al. (2014). Tempo and mode of regulatory evolution in *Drosophila*. *Genome Research* 24(5): 797-808.
doi: 10.1101/gr.163014.113

Cooper DW, Johnston PG, Watson JM, Graves JAM (1993). X-inactivation in marsupials and monotremes. *Seminars in Developmental Biology* 4(2): 117-128.
doi: 10.1006/sedb.1993.1014

Cortez D, et al. (2014). Origins and functional evolution of Y chromosomes across mammals. *Nature* 508(7497): 488-493.
doi: 10.1038/nature13151

Coulson AFW, Moult J (2002). A unifold, mesofold, and superfold model of protein fold use. *Proteins: Structure, Function, and Bioinformatics* 46(1): 61-71.
doi: 10.1002/prot.10011

Courchesne E, et al. (2001). Unusual brain growth patterns in early life in patients with autistic disorder an MRI study. *Neurology* 57(2): 245-254.
doi: -

Courchesne E, Pierce K (2005). Why the frontal cortex in autism might be talking only to itself: Local over-connectivity but long-distance disconnection. *Current Opinion in Neurobiology* 15(2): 225-230.
doi: doi:10.1016/j.conb.2005.03.001

Crow KD, Wagner GP (2006). What is the role of genome duplication in the evolution of complexity and diversity? *Molecular Biology and Evolution* 23(5): 887-892.
doi: 10.1093/molbev/msj083

Cui L, et al. (2006). Widespread genome duplications throughout the history of flowering plants. *Genome Research* 16(6): 738-749.
doi: 10.1101/gr.4825606

Cummins J. (2008). 5 Sperm motility and energetics. In. Sperm biology: An evolutionary perspective. p. 185.

Cunningham F, et al. (2015). Ensembl 2015. *Nucleic Acids Research* 43(D1): D662-D669.
doi: 10.1093/nar/gku1010

Davis JM, et al. (2014). DUF1220 dosage is linearly associated with increasing severity of the three primary symptoms of autism. *PLoS Genetics* 10(3): e1004241.
doi: 10.1371/journal.pgen.1004241

de Paula WBM, et al. (2013). Female and male gamete mitochondria are distinct and complementary in transcription, structure, and genome function. *Genome Biology and Evolution* 5(10): 1969-1977.
doi: 10.1093/gbe/evt147

Deakin J, Chaumeil J, Hore T, Marshall Graves J (2009). Unravelling the evolutionary origins of X chromosome inactivation in mammals: Insights from marsupials and monotremes. *Chromosome Research* 17(5): 671-685.
doi: 10.1007/s10577-009-9058-6

Dean R, et al. (2015). Positive selection underlies Faster-Z evolution of gene expression in birds. *Molecular Biology and Evolution* 32(10): 2646-2656.
doi: 10.1093/molbev/msv138

Dean R, Mank JE (2014). The role of sex chromosomes in sexual dimorphism: Discordance between molecular and phenotypic data. *Journal of Evolutionary Biology* 27(7): 1443-1453.
doi: 10.1111/jeb.12345

Dean R, Zimmer F, Mank JE (2014). The potential role of sexual conflict and sexual selection in shaping the genomic distribution of mito-nuclear genes. *Genome Biology and Evolution* 6(5): 1096-1104.
doi: 10.1093/gbe/evu063

Dehal P, Boore JL (2005). Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biology* 3(10): e314.
doi: 10.1371/journal.pbio.0030314

Deng X, et al. (2011). Evidence for compensatory upregulation of expressed X-linked genes in mammals, *Caenorhabditis elegans* and *Drosophila melanogaster*. *Nature Genetics* 43(12): 1179-1185.
doi: 10.1038/ng.948

Dennis MY, et al. (2012). Human-specific evolution of novel *SRGAP2* genes by incomplete segmental duplication. *Cell* 149(4): 912-922.
doi: 10.1016/j.cell.2012.03.033

Depino AM (2013). Peripheral and central inflammation in autism spectrum disorders. *Molecular and Cellular Neuroscience* 53(-): 69-76.
doi: 10.1016/j.mcn.2012.10.003

Des Marais DL, Rausher MD (2008). Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature* 454(7205): 762-765.
doi: 10.1038/nature07092

Deutschbauer AM, et al. (2005). Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. *Genetics* 169(4): 1915-1925.
doi: 10.1534/genetics.104.036871

Dey G, et al. (2013). Signaling network of Oncostatin M pathway. *Journal of Cell Communication and Signaling* 7(2): 103-108.
doi: 10.1007/s12079-012-0186-y

Dowling DK, Friberg U, Lindell J (2008). Evolutionary implications of non-neutral mitochondrial genetic variation. *Trends in Ecology & Evolution* 23(10): 546-554.
doi: 10.1016/j.tree.2008.05.011

Drăgan M-A, et al. (2016). GeneValidator: Identify problems with protein-coding gene predictions. *Bioinformatics* 32(10): 1559-1561.
doi: 10.1093/bioinformatics/btw015

Drown DM, Preuss KM, Wade MJ (2012). Evidence of a paucity of genes that interact with the mitochondrion on the X in mammals. *Genome Biology and Evolution* 4(8): 875-880.
doi: 10.1093/gbe/evs064

Dujon B (2010). Yeast evolutionary genomics. *Nature Reviews Genetics* 11(7): 512-524.
doi: 10.1038/nrg2811

Dumas L, Sikela JM (2009). DUF1220 domains, cognitive disease, and human brain evolution. *Cold Spring Harbor Symposia on Quantitative Biology* 74(-): 375-382.
doi: 10.1101/sqb.2009.74.025

Dumas LJ, et al. (2012). DUF1220-domain copy number implicated in human brain-size pathology and evolution. *The American Journal of Human Genetics* 91(3): 444-454.
doi: 10.1016/j.ajhg.2012.07.016

Dunn CW, Luo X, Wu Z (2013). Phylogenetic analysis of gene expression. *Integrative and Comparative Biology* 53(5): 847-856.
doi: 10.1093/icb/ict068

Durinck S, et al. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* 21(16): 3439-3440.
doi: 10.1093/bioinformatics/bti525

Dyall SD, Brown MT, Johnson PJ (2004). Ancient invasions: From endosymbionts to organelles. *Science* 304(5668): 253-257.
doi: 10.1126/science.1094884

Eddy SR (2011). Accelerated profile HMM searches. *PLoS Computational Biology* 7(10): e1002195.
doi: 10.1371/journal.pcbi.1002195

Edgar RC (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32(5): 1792-1797.
doi: 10.1093/nar/gkh340

Edger P, Pires JC (2009). Gene and genome duplications: The impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Research* 17(5): 699-717.
doi: 10.1007/s10577-009-9055-9

Ekblom R, Galindo J (2011). Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity* 107(1): 1-15.
doi: 10.1038/hdy.2010.152

Ellegren H (2005). The avian genome uncovered. *Trends in Ecology & Evolution* 20(4): 180-186.
doi: 10.1016/j.tree.2005.01.015

Ellegren H (2007). Characteristics, causes and evolutionary consequences of male-biased mutation. *Proceedings of the Royal Society of London B: Biological Sciences* 274(1606): 1-10.
doi: 10.1098/rspb.2006.3720

Ellegren H (2011). Sex-chromosome evolution: Recent progress and the influence of male and female heterogamety. *Nature Reviews Genetics* 12(3): 157-166.
doi: 10.1038/nrg2948

Ellegren H, et al. (2007). Faced with inequality: Chicken do not have a general dosage compensation of sex-linked genes. *BMC Biology* 5(1): 1-12.
doi: 10.1186/1741-7007-5-40

Emerson JJ, Kaessmann H, Betrán E, Long M (2004). Extensive gene traffic on the mammalian x chromosome. *Science* 303(5657): 537-540.
doi: 10.1126/science.1090042

Emerson RO, Thomas JH (2009). Adaptive evolution in zinc finger transcription factors. *PLoS Genetics* 5(1): e1000325.
doi: 10.1371/journal.pgen.1000325

Enright AJ, Van Dongen S, Ouzounis CA (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research* 30(7): 1575-1584.
doi: 10.1093/nar/30.7.1575

Ercan S (2015). Mechanisms of X chromosome dosage compensation. *Journal of Genomics* 3(-): 1-19.
doi: 10.7150/jgen.10404

Ermolaeva MD, Wu M, Eisen JA, Salzberg SL (2003). The age of the *Arabidopsis thaliana* genome duplication. *Plant Molecular Biology* 51(6): 859-866.
doi: 10.1023/A:1023001130337

Escriva H, et al. (2006). Neofunctionalization in vertebrates: The example of retinoic acid receptors. *PLoS Genetics* 2(7): e102.
doi: 10.1371/journal.pgen.0020102

Felsenstein J (1985). Phylogenies and the comparative method. *American Naturalist* 125(1): 1-15.
doi: -

Feng Y, et al. (2015). Nanopore-based fourth-generation DNA sequencing technology. *Genomics, Proteomics & Bioinformatics* 13(1): 4-16.
doi: 10.1016/j.gpb.2015.01.009

Feuk L, Carson AR, Scherer SW (2006). Structural variation in the human genome. *Nature Reviews Genetics* 7(2): 85-97.
doi: 10.1038/nrg1767

Finn RD, et al. (2014). Pfam: the protein families database. *Nucleic Acids Research* 42(D1): D222-D230.
doi: 10.1093/nar/gkt1223

Fitch WM (1970). Distinguishing homologous from analogous proteins. *Systematic Biology* 19(2): 99-113.
doi: 10.2307/2412448

Flicek P, et al. (2014). Ensembl 2014. *Nucleic Acids Research* 42(D1): D749-D755.
doi: 10.1093/nar/gkt1196

Force A, et al. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151(4): 1531-1545.
doi: -

Fortna A, et al. (2004). Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS Biology* 2(7): e207.
doi: 10.1371/journal.pbio.0020207

Francke U (1999). Williams-Beuren syndrome: Genes and mechanisms. *Human Molecular Genetics* 8(10): 1947-1954.
doi: 10.1093/hmg/8.10.1947

Frank SA, Hurst LD (1996). Mitochondria and male disease. *Nature* 383(6597): 224.
doi: 10.1038/383224a0

Frésard L, et al. (2014). Transcriptome-wide investigation of genomic imprinting in chicken. *Nucleic Acids Research* 42(6): 3768-3782.
doi: 10.1093/nar/gkt1390

Friedman R, Hughes AL (2001). Pattern and timing of gene duplication in animal genomes. *Genome Research* 11(11): 1842-1847.
doi: 10.1101/gr.200601

Gallach M, Chandrasekaran C, Betrán E (2010). Analyses of nuclearly encoded mitochondrial genes suggest gene duplication as a mechanism for resolving intralocus sexually antagonistic conflict in *Drosophila*. *Genome Biology and Evolution* 2(-): 835-850.
doi: 10.1093/gbe/evq069

Garland T, Jr. (2001). Phylogenetic comparison and artificial selection. In. Hypoxia: Springer. p. 107-132.

Gaut BS, Le Thierry d'Ennequin M, Peek AS, Sawkins MC (2000). Maize as a model for the evolution of plant nuclear genomes. *Proceedings of the National Academy of Sciences of the United States of America* 97(13): 7008-7015.
doi: 10.1073/pnas.97.13.7008

Gemmell NJ, Metcalf VJ, Allendorf FW (2004). Mother's curse: the effect of mtDNA on individual fitness and population viability. *Trends in Ecology & Evolution* 19(5): 238-244.
doi: 10.1016/j.tree.2004.02.002

Gillham NW. (1994). Organelle genes and genomes: Oxford University Press.

Gonzalez E, et al. (2005). The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* 307(5714): 1434-1440.
doi: 10.1126/science.1101160

Grabherr MG, et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* 29(7): 644-652.
doi: 10.1038/nbt.1883

Graves JAM (2006). Sex chromosome specialization and degeneration in mammals. *Cell* 124(5): 901-914.
doi: 10.1016/j.cell.2006.02.024

Graves JM (2014). Avian sex, sex chromosomes, and dosage compensation in the age of genomics. *Chromosome Research* 22(1): 45-57.
doi: 10.1007/s10577-014-9409-9

Haas BJ, et al. (2013). *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* 8(8): 1494-1512.
doi: 10.1038/nprot.2013.084

Hadfield JD (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software* 33(2): 1-22.
doi: -

Hahn MW (2009). Distinguishing among evolutionary models for the maintenance of gene duplicates. *Journal of Heredity* 100(5): 605-617.
doi: 10.1093/jhered/esp047

Hahn MW, Han MV, Han S-G (2007). Gene family evolution across 12 *Drosophila* genomes. *PLoS Genetics* 3(11): e197.
doi: 10.1371/journal.pgen.0030197

Haldane JBS (1933). The part played by recurrent mutation in evolution. *The American Naturalist* 67(708): 5-19.
doi: -

Harrison PW, et al. (2015). Sexual selection drives evolution and rapid turnover of male gene expression. *Proceedings of the National Academy of Sciences* 112(14): 4393-4398.
doi: 10.1073/pnas.1501339112

Hazlett HC, et al. (2005). Magnetic resonance imaging and head circumference study of brain size in autism: Birth through age 2 years. *Archives of General Psychiatry* 62(12): 1366-1376.
doi: 10.1001/archpsyc.62.12.1366

Hedges SB, et al. (2015). Tree of life reveals clock-like speciation and diversification. *Molecular Biology and Evolution* 32(4): 835-845.
doi: 10.1093/molbev/msv037

Hedrick PW (2012). Reversing mother's curse revisited. *Evolution* 66(2): 612-616.
doi: 10.1111/j.1558-5646.2011.01465.x

Hill GE (2014). Sex linkage of nuclear-encoded mitochondrial genes. *Heredity* 112(5): 469-470.
doi: 10.1038/hdy.2013.125

Hill GE, Johnson JD (2013). The mitonuclear compatibility hypothesis of sexual selection. *Proceedings of the Royal Society of London B: Biological Sciences* 280(1768): 20131314.
doi: 10.1098/rspb.2013.1314

Hillier LW, et al. (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432(7018): 695-716.
doi: 10.1038/nature03154

Hough J, Ågren JA, Barrett SCH, Wright SI (2014). Chromosomal distribution of cytonuclear genes in a dioecious plant with sex chromosomes. *Genome biology and evolution* 6(9): 2439-2443.
doi: 10.1093/gbe/evu197

Huerta-Cepas J, Dopazo J, Huynen MA, Gabaldón T (2011). Evidence for short-time divergence and long-time conservation of tissue-specific expression after gene duplication. *Briefings in Bioinformatics* 12(5): 442-448.
doi: 10.1093/bib/bbr022

Hughes AL (1994). The evolution of functionally novel proteins after gene duplication. *Proceedings of the Royal Society of London B: Biological Sciences* 256(1346): 119-124.
doi: 10.1098/rspb.1994.0058

Huminiecki L, Wolfe KH (2004). Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. *Genome Research* 14(10a): 1870-1879.
doi: 10.1101/gr.2705204

Hunter JD (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* 9(3): 90-95.
doi: 10.1109/MCSE.2007.55

Hurles M (2004). Gene duplication: The genomic trade in spare parts. *PLoS Biology* 2(7): e206.
doi: 10.1371/journal.pbio.0020206

Hurst LD, et al. (2015). The constrained maximal expression level owing to haploidy shapes gene content on the mammalian X chromosome. *PLoS Biology* 13(12): e1002315.
doi: 10.1371/journal.pbio.1002315

Inestrosa NC, Varela-Nallar L (2015). Wnt signalling in neuronal differentiation and development. *Cell and Tissue Research* 359(1): 215-223.
doi: 10.1007/s00441-014-1996-4

Innan H, Kondrashov F (2010). The evolution of gene duplications: Classifying and distinguishing between models. *Nature Reviews Genetics* 11(2): 97-108.
doi: 10.1038/nrg2689

Innocenti P, Morrow EH, Dowling DK (2011). Experimental evidence supports a sex-specific selective sieve in mitochondrial genome evolution. *Science* 332(6031): 845-848.
doi: 10.1126/science.1201157

Itoh Y, et al. (2007). Dosage compensation is less effective in birds than in mammals. *Journal of Biology* 6(1): 1-15.
doi: 10.1186/jbiol53

Ives AR (2015). For testing the significance of regression coefficients, go ahead and log-transform count data. *Methods in Ecology and Evolution* 6(7): 828-835.
doi: 10.1111/2041-210X.12386

Jacob F (1977). Evolution and tinkering. *Science* 196(4295): 1161-1166.
doi: 10.1126/science.860134

Jafari M, et al. (2013). Evolutionarily conserved motifs and modules in mitochondrial protein–protein interaction networks. *Mitochondrion* 13(6): 668-675.
doi: 10.1016/j.mito.2013.09.006

Jaillon O, et al. (2004). Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* 431(7011): 946-957.
doi: 10.1038/nature03025

Jarvis ED, et al. (2014). Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* 346(6215): 1320-1331.
doi: 10.1126/science.1253451

Jetz W, et al. (2012). The global diversity of birds in space and time. *Nature* 491(7424): 444-448.
doi: 10.1038/nature11631

Joseph SB, Kirkpatrick M (2004). Haploid selection in animals. *Trends in Ecology & Evolution* 19(11): 592-597.
doi: 10.1016/j.tree.2004.08.004

Julien P, et al. (2012). Mechanisms and evolutionary patterns of mammalian and avian dosage compensation. *PLoS Biology* 10(5): e1001328.
doi: 10.1371/journal.pbio.1001328

Kaessmann H, Vinckenbosch N, Long M (2009). RNA-based gene duplication: mechanistic and evolutionary insights. *Nature Reviews Genetics* 10(1): 19-31.
doi: 10.1038/nrg2487

Karin M, Lin A (2002). NF- κ B at the crossroads of life and death. *Nature Immunology* 3(3): 221-227.
doi: 10.1038/ni0302-221

Katoh K, Misawa K, Kuma Ki, Miyata T (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* 30(14): 3059-3066.
doi: 10.1093/nar/gkf436

Keeney JG, et al. (2015a). DUF1220 protein domains drive proliferation in human neural stem cells and are associated with increased cortical volume in anthropoid primates. *Brain Structure and Function* 220(5): 3053-3060.
doi: 10.1007/s00429-014-0814-9

Keeney JG, Dumas L, Sikela JM (2014). The case for DUF1220 domain dosage as a primary contributor to anthropoid brain expansion. *Frontiers in Human Neuroscience* 8(-): 427.
doi: 10.3389/fnhum.2014.00427

Keeney JG, et al. (2015b). Generation of mice lacking DUF1220 protein domains: effects on fecundity and hyperactivity. *Mammalian Genome* 26(1-2): 33-42.
doi: 10.1007/s00335-014-9545-8

Kellis M, Birren BW, Lander ES (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* 428(6983): 617-624.
doi: 10.1038/nature02424

Kersey PJ, et al. (2012). Ensembl Genomes: An integrative resource for genome-scale data from non-vertebrate species. *Nucleic Acids Research* 40(D1): D91-D97.
doi: 10.1093/nar/gkr895

Khil PP, Smirnova NA, Romanienko PJ, Camerini-Otero RD (2004). The mouse X chromosome is enriched for sex-biased genes not subject to selection by meiotic sex chromosome inactivation. *Nature Genetics* 36(6): 642-646.
doi: 10.1038/ng1368

Kidwell MC, et al. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLoS Biology* 14(5): e1002456.
doi: 10.1371/journal.pbio.1002456

Kim D, et al. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* 14(4): R36.
doi: 10.1186/gb-2013-14-4-r36

Kimura M. (1984). The neutral theory of molecular evolution: Cambridge University Press.

Kimura M, King JL (1979). Fixation of a deleterious allele at one of two "duplicate" loci by mutation pressure and random drift. *Proceedings of the National Academy of Sciences of the United States of America* 76(6): 2858-2861.
doi: -

Kitano J, et al. (2009). A role for a neo-sex chromosome in stickleback speciation. *Nature* 461(7267): 1079-1083.
doi: 10.1038/nature08441

Koboldt DC, et al. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research* 22(3): 568-576.
doi: 10.1101/gr.129684.111

Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV (2002). Selection in the evolution of gene duplications. *Genome Biology* 3(2): research0008.0001-research0008.0009.
doi: -

Koonin EV (2005). Orthologs, paralogs, and evolutionary genomics. *Annual Review of Genetics* 39(-):309-338.
doi: 10.1146/annurev.genet.39.073003.114725

Kuan C-Y, Roth KA, Flavell RA, Rakic P (2000). Mechanisms of programmed cell death in the developing brain. *Trends in Neurosciences* 23(7): 291-297.
doi: 10.1016/S0166-2236(00)01581-2

Lahn BT, Page DC (1999). Four evolutionary strata on the human X chromosome. *Science* 286(5441): 964-967.
doi: 10.1126/science.286.5441.964

Lander ES, et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409(6822): 860-921.
doi: 10.1038/35057062

Langmead B, Salzberg SL (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9(4): 357-359.
doi: 10.1038/nmeth.1923

Larhammar D, Lundin L-G, Hallböök F (2002). The human hox-bearing chromosome regions did arise by block or chromosome (or even genome) duplications. *Genome Research* 12(12): 1910-1920.
doi: 10.1101/gr.445702

Lau AN, et al. (2009). Horse domestication and conservation genetics of Przewalski's horse inferred from sex chromosomal and autosomal sequences. *Molecular Biology and Evolution* 26(1): 199-208.
doi: 10.1093/molbev/msn239

Leder EH, et al. (2010). Female-biased expression on the X chromosome as a key step in sex chromosome evolution in threespine sticklebacks. *Molecular Biology and Evolution* 27(7): 1495-1503.
doi: 10.1093/molbev/msq031

Lejeune J, Gautier M, Turpin R (1959). Les chromosomes humains en culture de tissus. *Comptes Rendus Academies des Sciences* 248(-):602–603.
doi: -

Lejeune JG, Gautier M, Turpin R (1957). Etude des chromosomes somatiques de neuf enfants mongoliens. *Comptes Rendus Academies des Sciences* 248(-):1721.
doi: -

Lemons D, McGinnis W (2006). Genomic evolution of hox gene clusters. *Science* 313(5795): 1918-1922.
doi: 10.1126/science.1132040

Leprevost FdV, da, et al. (2014). On best practices in the development of bioinformatics software. *Frontiers in Genetics* 5(-): 199.
doi: 10.3389/fgene.2014.00199

Levitt M (2009). Nature of the protein universe. *Proceedings of the National Academy of Sciences of the United States of America* 106(27): 11079-11084.
doi: 10.1073/pnas.0905029106

Li B, Dewey CN (2011). RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12(1): 1-16.
doi: 10.1186/1471-2105-12-323

Li H, et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16): 2078-2079.
doi: 10.1093/bioinformatics/btp352

Lin F, Xing K, Zhang J, He X (2012). Expression reduction in mammalian X chromosome evolution refutes Ohno's hypothesis of dosage compensation. *Proceedings of the National Academy of Sciences of the United States of America* 109(29): 11752-11757.
doi: 10.1073/pnas.1201816109

Lin Y, et al. (2016). Comparison of normalization and differential expression analyses using RNA-Seq data from 726 individual *Drosophila melanogaster*. *BMC Genomics* 17(1): 1-20.
doi: 10.1186/s12864-015-2353-z

Lipinski KJ, et al. (2011). High spontaneous rate of gene duplication in *Caenorhabditis elegans*. *Current Biology* 21(4): 306-310.
doi: 10.1016/j.cub.2011.01.026deha

Liu Y, Zhou J, White KP (2014). RNA-seq differential expression studies: More sequence or more replication? *Bioinformatics* 30(3): 301-304.
doi: 10.1093/bioinformatics/btt688

Livernois AM, et al. (2013). Independent evolution of transcriptional inactivation on sex chromosomes in birds and mammals. *PLoS Genetics* 9(7): e1003635.
doi: 10.1371/journal.pgen.1003635

Lohse M, et al. (2012). RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Research* 40(W1): W622-W627.
doi: 10.1093/nar/gks540

Lotz C, et al. (2013). Characterization, design, and function of the mitochondrial proteome: From organs to organisms. *Journal of Proteome Research* 13(2): 433-446.
doi: 10.1021/pr400539j

Lynch M, Conery JS (2000). The evolutionary fate and consequences of duplicate genes. *Science* 290(5494): 1151-1155.
doi: 10.1126/science.290.5494.1151

Lynch M, Force A (2000). The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154(1): 459-473.
doi: -

Madden SD, Donovan M, Cotter TG (2007). Key apoptosis regulating proteins are down-regulated during postnatal tissue development. *International Journal of Developmental Biology* 51(5): 415.
doi: 10.1387/ijdb.062263sm

Makałowski W (2001). Are we polyploids? A brief history of one hypothesis. *Genome Research* 11(5): 667-670.
doi: 10.1101/gr.188801

Makino T, Hokamp K, McLysaght A (2009). The complex relationship of gene duplication and essentiality. *Trends in Genetics* 25(4): 152-155.
doi: 10.1016/j.tig.2009.03.001

Makino T, McLysaght A (2010). Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proceedings of the National Academy of Sciences of the United States of America* 107(20): 9270-9274.
doi: 10.1073/pnas.0914697107

Makino T, McLysaght A, Kawata M (2013). Genome-wide deserts for copy number variation in vertebrates. *Nature Communications* 4(-):2283.
doi: 10.1038/ncomms3283

Makova KD, Li W-H (2003). Divergence in the spatial pattern of gene expression between human duplicate genes. *Genome Research* 13(7): 1638-1645.
doi: 10.1101/gr.1133803

Malone JH, et al. (2012). Mediation of *Drosophila* autosomal dosage effects and compensation by network interactions. *Genome Biology* 13(4): R28-R28.
doi: 10.1186/gb-2012-13-4-r28

Mank JE (2013). Sex chromosome dosage compensation: Definitely not for everyone. *Trends in Genetics* 29(12): 677-683.
doi: 10.1016/j.tig.2013.07.005

Mank JE, Ellegren H (2008). All dosage compensation is local: Gene-by-gene regulation of sex-biased expression on the chicken Z chromosome. *Heredity* 102(3): 312-320.
doi: 10.1038/hdy.2008.116

Mank JE, Hosken DJ, Wedell N (2011). Some inconvenient truths about sex chromosome dosage compensation and the potential role of sexual conflict. *Evolution* 65(8): 2133-2144.
doi: 10.1111/j.1558-5646.2011.01316.x

Mank JE, Vicoso B, Berlin S, Charlesworth B (2010). Effective population size and the Faster-X effect: empirical results and their interpretation. *Evolution* 64(3): 663-674.
doi: 10.1111/j.1558-5646.2009.00853.x

Marcet-Houben M, Gabaldón T (2015). Beyond the whole-genome duplication: phylogenetic evidence for an ancient interspecies hybridization in the Baker's yeast lineage. *PLoS Biology* 13(8): e1002220.
doi: 10.1371/journal.pbio.1002220

McLysaght A, Hokamp K, Wolfe KH (2002). Extensive genomic duplication during early chordate evolution. *Nature Genetics* 31(2): 200-204.
doi: 10.1038/ng884

McLysaght A, et al. (2014). Ohnologs are overrepresented in pathogenic copy number mutations. *Proceedings of the National Academy of Sciences of the United States of America* 111(1): 361-366.
doi: 10.1073/pnas.1309324111

McPeek Mark A, Brown Jonathan M (2007). Clade age and not diversification rate explains species richness among animal taxa. *The American Naturalist* 169(4): E97-E106.
doi: 10.1086/512135

Meiklejohn CD, et al. (2013). An incompatibility between a mitochondrial tRNA and its nuclear-encoded tRNA synthetase compromises development and fitness in *Drosophila*. *PLoS Genetics* 9(1): e1003238.
doi: 10.1371/journal.pgen.1003238

Meisel RP, Han MV, Hahn MW (2009). A complex suite of forces drives gene traffic from *Drosophila* X chromosomes. *Genome Biology and Evolution* 1(-): 176-188.
doi: 10.1093/gbe/evp018

Meisel RP, Malone JH, Clark AG (2012). Disentangling the relationship between sex-biased gene expression and X-linkage. *Genome Research* 22(7): 1255-1265.
doi: 10.1101/gr.132100.111

Meyer BJ (2010). Targeting X chromosomes for repression. *Current Opinion in Genetics & Development* 20(2): 179-189.
doi: 10.1016/j.gde.2010.03.008

Meyer U, Feldon J, Dammann O (2011). Schizophrenia and autism: Both shared and disorder-specific pathogenesis via perinatal inflammation? *Pediatric Research* 69(26R-33R.
doi: 10.1203/PDR.0b013e318212c196

Miller JR, Koren S, Sutton G (2010). Assembly algorithms for next-generation sequencing data. *Genomics* 95(6): 315-327.
doi: 10.1016/j.ygeno.2010.03.001

Minshew NJ, Williams DL (2007). The new neurobiology of autism: Cortex, connectivity, and neuronal organization. *Archives of Neurology* 64(7): 945-950.
doi: 10.1001/archneur.64.7.945

Moghadam HK, et al. (2012). W chromosome expression responds to female-specific selection. *Proceedings of the National Academy of Sciences of the United States of America* 109(21): 8207-8211.
doi: 10.1073/pnas.1202721109

Montgomery S, Mank J (2016). Inferring regulatory change from gene expression: the confounding effects of tissue scaling. *In Review.*

Montgomery SH, et al. (2011). Adaptive evolution of four microcephaly genes and the evolution of brain size in anthropoid primates. *Molecular Biology and Evolution* 28(1): 625-638.
doi: 10.1093/molbev/msq237

Montgomery SH, Mundy NI (2012a). Evolution of *ASPM* is associated with both increases and decreases in brain size in primates. *Evolution* 66(3): 927-932.
doi: 10.1111/j.1558-5646.2011.01487.x

Montgomery SH, Mundy NI (2012b). Positive selection on *NIN*, a gene involved in neurogenesis, and primate brain evolution. *Genes, Brain and Behavior* 11(8): 903-910.
doi: 10.1111/j.1601-183X.2012.00844.x

Montooth KL, Meiklejohn CD, Abt DN, Rand DM (2010). Mitochondrial–nuclear epistasis affects fitness within species but does not contribute to fixed incompatibilities between species of *Drosophila*. *Evolution* 64(12): 3364-3379.
doi: 10.1111/j.1558-5646.2010.01077.x

Montooth KL, Rand DM (2008). The spectrum of mitochondrial mutation differs across species. *PLoS Biology* 6(8): e213.
doi: 10.1371/journal.pbio.0060213

Moore AD, et al. (2008). Arrangements in the modular evolution of proteins. *Trends in Biochemical Sciences* 33(9): 444-451.
doi: 10.1016/j.tibs.2008.05.008

Muller HJ (1914). A gene for the fourth chromosome of *Drosophila*. *Journal of Experimental Zoology* 17(3): 325-336.
doi: 10.1002/jez.1400170303

Muller HJ (1935). The origination of chromatin deficiencies as minute deletions subject to insertion elsewhere. *Genetica* 17(-): 237-252.
doi: 10.1007/BF01985012

Murphy WJ, et al. (1999). Extensive conservation of sex chromosome organization between cat and human revealed by parallel radiation hybrid mapping. *Genome Research* 9(12): 1223-1230.
doi: -

Natri HM, Shikano T, Merilä J (2013). Progressive recombination suppression and differentiation in recently evolved neo-sex chromosomes. *Molecular Biology and Evolution* 30(5): 1131-1144.
doi: 10.1093/molbev/mst035

Naurin S, et al. (2011). The sex-biased brain: sexual dimorphism in gene expression in two species of songbirds. *BMC Genomics* 12(1): 1-11.
doi: 10.1186/1471-2164-12-37

Needleman SB, Wunsch CD (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* 48(3): 443-453.
doi: 10.1016/0022-2836(70)90057-4

Nei M (1969). Gene duplication and nucleotide substitution in evolution. *Nature* 221(5175): 40-42.
doi: 10.1038/221040a0

Nguyen Ba AN, et al. (2014). Detecting functional divergence after gene duplication through evolutionary changes in posttranslational regulatory sequences. *PLoS Computational Biology* 10(12): e1003977.
doi: 10.1371/journal.pcbi.1003977

Nguyen D-Q, Webber C, Ponting CP (2006). Bias of selection on human copy-number variants. *PLoS Genetics* 2(2): e20.
doi: 10.1371/journal.pgen.0020020

O'Bleness M, et al. (2014). Finished sequence and assembly of the DUF1220-rich 1q21 region using a haploid human genome. *BMC Genomics* 15(1): 1.
doi: 10.1186/1471-2164-15-387

O'Bleness MS, et al. (2012). Evolutionary history and genome organization of DUF1220 protein domains. *G3: Genes | Genomes | Genetics* 2(9): 977-986.
doi: 10.1534/g3.112.003061

O'hara RB, Kotze DJ (2010). Do not log-transform count data. *Methods in Ecology and Evolution* 1(2): 118-122.
doi: 10.1111/j.2041-210X.2010.00021.x

Ohno S. (1970). Evolution by gene duplication: Springer Verlag.

Ohno S. (1967). Sex chromosomes and sex-linked genes: Springer Berlin Heidelberg.

Ohno S, Kaplan WD, Kinosita R (1959). Formation of the sex chromatin by a single X-chromosome in liver cells of *Rattus norvegicus*. *Experimental Cell Research* 18(2): 415-418.
doi: 10.1016/0014-4827(59)90031-X

Ohno S, Wolf U, Atkin NB (1968). Evolution from fish to mammals by gene duplication. *Hereditas* 59(1): 169-187.
doi: 10.1111/j.1601-5223.1968.tb02169.x

Organ CL, et al. (2007). Origin of avian genome size and structure in non-avian dinosaurs. *Nature* 446(7132): 180-184.
doi: 10.1038/nature05621

Pagel M (1999). Inferring the historical patterns of biological evolution. *Nature* 401(6756): 877-884.
doi: 10.1038/44766

Papp B, Pal C, Hurst LD (2003). Dosage sensitivity and the evolution of gene families in yeast. *Nature* 424(6945): 194-197.
doi: 10.1038/nature01771

Partridge L, Hurst LD (1998). Sex and conflict. *Science* 281(5385): 2003-2008.
doi: 10.1126/science.281.5385.2003

Paterson AH, Bowers JE, Chapman BA (2004). Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proceedings of the National Academy of Sciences of the United States of America* 101(26): 9903-9908.
doi: -

Payer B, Lee JT (2008). X chromosome dosage compensation: How mammals keep the balance. *Annual Review of Genetics* 42(1): 733-772.
doi: 10.1146/annurev.genet.42.110807.091711

Pérez F, Granger BE (2007). IPython: A system for interactive scientific computing. *Computing in Science & Engineering* 9(3): 21-29.
doi: 10.1109/MCSE.2007.53

Perkins ND (2012). The diverse and complex roles of NF- κ B subunits in cancer. *Nature Reviews Cancer* 12(2): 121-132.
doi: 10.1038/nrc3204

Perry GH, et al. (2007). Diet and the evolution of human amylase gene copy number variation. *Nature Genetics* 39(10): 1256-1260.
doi: 10.1038/ng2123

Pessia E, Engelstädter J, Marais GAB (2013). The evolution of X chromosome inactivation in mammals: the demise of Ohno's hypothesis? *Cellular and Molecular Life Sciences* 71(8): 1383-1394.
doi: 10.1007/s00018-013-1499-6

Pessia E, et al. (2012). Mammalian X chromosome inactivation evolved as a dosage-compensation mechanism for dosage-sensitive genes on the X chromosome. *Proceedings of the National Academy of Sciences of the United States of America* 109(14): 5346-5351.
doi: 10.1073/pnas.1116763109

Piatigorsky J (1991). The recruitment of crystallins: New functions precede gene duplication. *Science* 252(5009): 1078-1079.
doi: 10.1126/science.252.5009.1078

Pollack JR, et al. (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proceedings of the National Academy of Sciences of the United States of America* 99(20): 12963-12968.
doi: 10.1073/pnas.162471999

Polster BM, et al. (2003). Postnatal brain development and neural cell differentiation modulate mitochondrial Bax and BH3 peptide-induced cytochrome c release. *Cell Death & Differentiation* 10(3): 365-370.
doi: 10.1038/sj.cdd.4401158

Popesco MC, et al. (2006). Human lineage–specific amplification, selection, and neuronal expression of DUF1220 domains. *Science* 313(5791): 1304-1307.
doi: 10.1126/science.1127980

Popovici C, Leveugle M, Birnbaum D, Coulier F (2001). Coparalogy: Physical and functional clusterings in the human genome. *Biochemical and Biophysical Research Communications* 288(2): 362-370.
doi: 10.1006/bbrc.2001.5794

Porcelli D, Barsanti P, Pesole G, Caggese C (2007). The nuclear OXPHOS genes in insecta: A common evolutionary origin, a common *cis*-regulatory motif, a common destiny for gene duplicates. *BMC Evolutionary Biology* 7(215): -.
doi: 10.1186/1471-2148-7-215.

Potrzebowski L, et al. (2008). Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. *PLoS Biology* 6(4): e80.
doi: 10.1371/journal.pbio.0060080

Poulton CJ, et al. (2011). Microcephaly with simplified gyration, epilepsy, and infantile diabetes linked to inappropriate apoptosis of neural progenitors. *The American Journal of Human Genetics* 89(2): 265-276.
doi: 10.1016/j.ajhg.2011.07.006

Price RN, et al. (2004). Mefloquine resistance in *Plasmodium falciparum* and increased *pfmdr1* gene copy number. *The Lancet* 364(9432): 438-447.
doi: 10.1016/S0140-6736(04)16767-6

Quinlan AR, Hall IM (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6): 841-842.
doi: 10.1093/bioinformatics/btq033

Quinn A, Juneja P, Jiggins FM (2014a). Estimates of allele-specific expression in *Drosophila* with a single genome sequence and RNA-seq data. *Bioinformatics* 30(18): 2603-2610.
doi: 10.1093/bioinformatics/btu342

Quinn AM, Juneja P, Jiggins FM. 2014b. http://doi.org/10.5281/zenodo.10289.

R Core Team. 2015. R: A language and environment for statistical computing: R Foundation for Statistical Computing.

Raj A, van Oudenaarden A (2008). Stochastic gene expression and its consequences. *Cell* 135(2): 216-226.
doi: 10.1016/j.cell.2008.09.050

Rand DM, Clark AG, Kann LM (2001). Sexually antagonistic cytonuclear fitness interactions in *Drosophila melanogaster*. *Genetics* 159(1): 173-187.
doi: -

Rand DM, Haney RA, Fry AJ (2004). Cytonuclear coevolution: the genomics of cooperation. *Trends in Ecology & Evolution* 19(12): 645-653.
doi: 10.1016/j.tree.2004.10.003

Raudsepp T, et al. (2004). Exceptional conservation of horse–human gene order on X chromosome revealed by high-resolution radiation hybrid mapping. *Proceedings of the National Academy of Sciences of the United States of America* 101(8): 2386-2391.
doi: 10.1073/pnas.0308513100

Redon R, et al. (2006). Global variation in copy number in the human genome. *Nature* 444(7118): 444-454.
doi: 10.1038/nature05329

Reimand J, Arak T, Vilo J (2011). g:Profiler—a web server for functional interpretation of gene lists (2011 update). *Nucleic Acids Research* 39(suppl 2): W307-W315.
doi: 10.1093/nar/gkr378

Rhind N, et al. (2011). Comparative functional genomics of the fission yeasts. *Science* 332(6032): 930-936.
doi: 10.1126/science.1203357

Rice P, Longden I, Bleasby A (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* 16(6): 276-277.
doi: 10.1016/S0168-9525(00)02024-2

Rice WR (1984). Sex chromosomes and the evolution of sexual dimorphism. *Evolution* 38(4): 735-742.
doi: 10.2307/2408385

Robinson MD, McCarthy DJ, Smyth GK (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26(1): 139-140.
doi: 10.1093/bioinformatics/btp616

Robinson MD, Oshlack A (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* 11(3): R25.
doi: 10.1186/gb-2010-11-3-r25

Robles JA, et al. (2012). Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. *BMC Genomics* 13(1): 1-14.
doi: 10.1186/1471-2164-13-484

Rogell B, Dean R, Lemos B, Dowling DK (2014). Mito-nuclear interactions as drivers of gene movement on and off the X-chromosome. *BMC Genomics* 15(1): 1.
doi: 10.1186/1471-2164-15-330

Rogozin IB, Managadze D, Shabalina SA, Koonin EV (2014). Gene family level comparative analysis of gene expression in mammals validates the ortholog conjecture. *Genome Biology and Evolution* 6(4): 754-762.
doi: 10.1093/gbe/evu051

Rossmann MG, Moras D, Olsen KW (1974). Chemical and biological evolution of a nucleotide-binding protein. *Nature* 250(5463): 194-199.
doi: 10.1038/250194a0

Roth ACJ, Gonnet GH, Dessimoz C (2008). Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics* 9(1): 1-10.
doi: 10.1186/1471-2105-9-518

Roubin R, et al. (2013). Myomegalin is necessary for the formation of centrosomal and Golgi-derived microtubules. *Biology Open* 2(2): 238-250.
doi: 10.1242/bio.20123392

Rozowsky J, et al. (2011). AlleleSeq: Analysis of allele-specific expression and binding in a network framework. *Molecular Systems Biology* 7(1): 522.
doi: 10.1038/msb.2011.54

Ruepp A, et al. (2010). CORUM: the comprehensive resource of mammalian protein complexes—2009. *Nucleic Acids Research* 38(Database issue): D497-D501.
doi: 10.1093/nar/gkp914

Saitou N, Nei M (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4(4): 406-425.
doi: -

Sanno H, et al. (2010). Control of postnatal apoptosis in the neocortex by RhoA-subfamily GTPases determines neuronal density. *The Journal of Neuroscience* 30(12): 4221-4231.
doi: 10.1523/JNEUROSCI.3318-09.2010

Sato Y, Nishida M (2010). Teleost fish with specific genome duplication as unique models of vertebrate evolution. *Environmental Biology of Fishes* 88(2): 169-188.
doi: 10.1007/s10641-010-9628-7

Scally A, et al. (2012). Insights into hominid evolution from the gorilla genome sequence. *Nature* 483(7388): 169-175.
doi: 10.1038/nature10842

Scannell DR, et al. (2006). Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* 440(7082): 341-345.
doi: 10.1038/nature04562

Schmitz JF, Zimmer F, Bornberg-Bauer E (2016). Mechanisms of transcription factor evolution in Metazoa. *Nucleic Acids Research* In press
doi:10.1093/nar/gkw492

Schnoes AM, et al. (2013). Biases in the experimental annotations of protein function and their effect on our understanding of protein function space. *PLoS Computational Biology* 9(5): e1003063.
doi: 10.1371/journal.pcbi.1003063

Schwander T, Libbrecht R, Keller L (2014). Supergenes and complex phenotypes. *Current Biology* 24(7): R288-R294.
doi: 10.1016/j.cub.2014.01.056

Seoighe C, Gehring C (2004). Genome duplication led to highly selective expansion of the *Arabidopsis thaliana* proteome. *Trends in Genetics* 20(10): 461-464.
doi: 10.1016/j.tig.2004.07.008

Shimomura M, et al. (2009). KAIKObase: an integrated silkworm genome database and data mining tool. *BMC Genomics* 10(1): 486.
doi: 0.1186/1471-2164-10-486

Shlien A, Malkin D (2009). Copy number variations and cancer. *Genome Medicine* 1(6): 62-62.
doi: 10.1186/gm62

Singh ND, Macpherson JM, Jensen JD, Petrov DA (2007). Similar levels of X-linked and autosomal nucleotide variation in African and non-African populations of *Drosophila melanogaster*. *BMC Evolutionary Biology* 7(1): 1.
doi: 10.1186/1471-2148-7-202

Singh PP, Arora J, Isambert H (2015). Identification of ohnolog genes originating from whole genome duplication in early vertebrates, based on synteny comparison across multiple genomes. *PLoS Computational Biology* 11(7): e1004394.
doi: 10.1371/journal.pcbi.1004394

Smith S, Turbill C, Suchentrunk F (2010). Introducing mother's curse: Low male fertility associated with an imported mtDNA haplotype in a captive colony of brown hares. *Molecular Ecology* 19(1): 36-43.
doi: 10.1111/j.1365-294X.2009.04444.x

Smith TF, Waterman MS (1981). Identification of common molecular subsequences. *Journal of Molecular Biology* 147(1):195-197.
doi: 10.1016/0022-2836(81)90087-5

Soltis DE, et al. (2009). Polyploidy and angiosperm diversification. *American Journal of Botany* 96(1): 336-348.
doi: 10.3732/ajb.0800079

Soltis PS, Marchant DB, Van de Peer Y, Soltis DE (2015). Polyploidy and genome evolution in plants. *Current Opinion in Genetics & Development* 35(119-125.
doi: 10.1016/j.gde.2015.11.003

Sonnhammer ELL, Koonin EV (2002). Orthology, paralogy and proposed classification for paralog subtypes. *Trends in Genetics* 18(12): 619-620.
doi: 10.1016/S0168-9525(02)02793-2

Sonnhammer ELL, Östlund G (2015). InParanoid 8: Orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Research* 43(D1): D234-D239.
doi: 10.1093/nar/gku1203

Stephan H, Frahm H, Baron G. (1981). New and revised data on volumes of brain structures in insectivores and primates. In. Folia Primatologica. 35. p. 1-29.

Stevenson K, Coolon J, Wittkopp P (2013). Sources of bias in measures of allele-specific expression derived from RNA-seq data aligned to a single reference genome. *BMC Genomics* 14(1): 536.
doi: 10.1186/1471-2164-14-536

Stoltzfus A (1999). On the possibility of constructive neutral evolution. *Journal of Molecular Evolution* 49(2): 169-181.
doi: 10.1007/PL00006540

Straub T, Becker PB (2007). Dosage compensation: The beginning and end of generalization. *Nature Reviews Genetics* 8(1): 47-57.

Sturtevant AH (1925). The effects of unequal crossing over at the bar locus in *Drosophila*. *Genetics* 10(2): 117.
doi: -

Sturtevant AH (1913). The linear arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of association. *Journal of Experimental Zoology* 14(1): 43-59.
doi: 10.1002/jez.1400140104

Sugino RP, Innan H (2005). Estimating the time to the whole-genome duplication and the duration of concerted evolution via gene conversion in yeast. *Genetics* 171(1): 63-69.
doi: 10.1534/genetics.105.043869

Sundström H, Webster MT, Ellegren H (2004). Reduced variation on the chicken Z chromosome. *Genetics* 167(1): 377-385.
doi: 10.1534/genetics.167.1.377

Tak PP, Firestein GS (2001). NF- κ B: A key role in inflammatory diseases. *The Journal of Clinical Investigation* 107(1): 7-11.
doi: 10.1172/JCI11830

Taylor JS, Raes J (2004). Duplication and divergence: The evolution of new genes and old ideas. *Annual Review of Genetics* 38(1): 615-643.
doi: 10.1146/annurev.genet.38.072902.092831

Taylor RW, Turnbull DM (2005). Mitochondrial DNA mutations in human disease. *Nature Reviews Genetics* 6(5): 389-402.
doi: 10.1038/nrg1606

Terrapon N, Gascuel O, Maréchal É, Bréhélin L (2012). Fitting Hidden Markov Models of protein domains to a target species: application to *Plasmodium falciparum*. *BMC Bioinformatics* 13(1): 1-14.
doi: 10.1186/1471-2105-13-67

The Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408(6814): 796-815.
doi: 10.1038/35048692

Thornton GK, Woods CG (2009). Primary microcephaly: Do all roads lead to Rome? *Trends in Genetics* 25(11): 501-510.
doi: 10.1016/j.tig.2009.09.011

Timmis JN, Ayliffe MA, Huang CY, Martin W (2004). Endosymbiotic gene transfer: Organelle genomes forge eukaryotic chromosomes. *Nature Reviews Genetics* 5(2): 123-135.
doi: 10.1038/nrg1271

Torres EM, et al. (2007). Effects of aneuploidy on cellular physiology and cell division in haploid yeast. *Science* 317(5840): 916-924.
doi: 10.1126/science.1142210

Toups MA, Pease JB, Hahn MW (2011). No excess gene movement is detected off the avian or lepidopteran Z chromosome. *Genome Biology and Evolution* 3(-): 1381-1390.
doi: 10.1093/gbe/evr109

Turner HH (1938). A syndrome of infantilism, congenital webbed neck, and cubitus valgus. *Endocrinology* 23(5): 566-574.
doi: 10.1210/endo-23-5-566

Uebbing S, et al. (2015). Quantitative mass spectrometry reveals partial translational regulation for dosage compensation in chicken. *Molecular Biology and Evolution* 32(10): 2716-2725.
doi: 10.1093/molbev/msv147

Uebbing S, Künstner A, Mäkinen H, Ellegren H (2013). Transcriptome sequencing reveals the character of incomplete dosage compensation across multiple tissues in flycatchers. *Genome Biology and Evolution* 5(8): 1555-1566.
doi: 10.1093/gbe/evt114

Unckless RL, Herren JK (2009). Population genetics of sexually antagonistic mitochondrial mutants under inbreeding. *Journal of Theoretical Biology* 260(1): 132-136.
doi: 10.1016/j.jtbi.2009.06.004

Uno Y, et al. (2013). Homoeologous chromosomes of *Xenopus laevis* are highly conserved after whole-genome duplication. *Heredity* 111(5): 430-436.
doi: 10.1038/hdy.2013.65

Vallender EJ, Lahn BT (2006). A primate-specific acceleration in the evolution of the caspase-dependent apoptosis pathway. *Human Molecular Genetics* 15(20): 3034-3040.
doi: 10.1093/hmg/ddl245

Van de Peer Y, Maere S, Meyer A (2009). The evolutionary significance of ancient genome duplications. *Nature Reviews Genetics* 10(10): 725-732.
doi: 10.1038/nrg2600

van der Heijden RTJM, Snel B, van Noort V, Huynen MA (2007). Orthology prediction at scalable resolution by phylogenetic tree analysis. *BMC Bioinformatics* 8(1): 1-12.
doi: 10.1186/1471-2105-8-83

Vandepoele K, et al. (2008). A constitutional translocation t (1; 17)(p36. 2; q11. 2) in a neuroblastoma patient disrupts the human *NBPF1* and *ACCN1* genes. *PloS One* 3(5): e2207.
doi: 10.1371/journal.pone.0002207

Vandepoele K, Staes K, Andries V, Van Roy F (2010). Chibby interacts with *NBPF1* and clusterin, two candidate tumor suppressors linked to neuroblastoma. *Experimental Cell Research* 316(7): 1225-1233.
doi: 10.1016/j.yexcr.2010.01.019

Vandepoele K, et al. (2005). A novel gene family NBPF: intricate structure generated by gene duplications during primate evolution. *Molecular Biology and Evolution* 22(11): 2265-2274.
doi: 10.1093/molbev/msi222

Veitia RA (2002). Exploring the etiology of haploinsufficiency. *BioEssays* 24(2): 175-184.
doi: 10.1002/bies.10023

Veitia RA (2004). Gene dosage balance in cellular pathways: implications for dominance and gene duplicability. *Genetics* 168(1): 569-574.
doi: 10.1534/genetics.104.029785

Veitia RA, Veyrunes F, Bottani S, Birchler JA (2015). X chromosome inactivation and active X upregulation in therian mammals: facts, questions, and hypotheses. *Journal of Molecular Cell Biology* 7(1): 2-11.
doi: 10.1093/jmcb/mjv001

Venditti C, Meade A, Pagel M (2011). Multiple routes to mammalian diversity. *Nature* 479(7373): 393-396.
doi: 10.1038/nature10516

Venter JC, et al. (2001). The sequence of the human genome. *Science* 291(5507): 1304-1351.
doi: 10.1126/science.1058040

Verstrepen KJ, Jansen A, Lewitter F, Fink GR (2005). Intragenic tandem repeats generate functional variability. *Nature Genetics* 37(9): 986-990.
doi: 10.1038/ng1618

Veyrunes F, et al. (2008). Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes. *Genome Research* 18(6): 965-973.
doi: 10.1101/gr.7101908

Vibranovski MD, Zhang Y, Long M (2009). General gene movement off the X chromosome in the *Drosophila* genus. *Genome Research* 19(5): 897-903.
doi: 10.1101/gr.088609.108

Vicoso B, Bachtrog D (2011). Lack of global dosage compensation in *Schistosoma mansoni*, a female-heterogametic parasite. *Genome Biology and Evolution* 3(-):230-235.
doi: 10.1093/gbe/evr010

Vicoso B, Bachtrog D (2015). Numerous transitions of sex chromosomes in *Diptera*. *PLoS Biology* 13(4): e1002078.
doi: 10.1371/journal.pbio.1002078

Vicoso B, Bachtrog D (2013). Reversal of an ancient sex chromosome to an autosome in *Drosophila*. *Nature* 499(7458): 332-335.

Vicoso B, et al. (2013). Comparative sex chromosome genomics in snakes: differentiation, evolutionary strata, and lack of global dosage compensation. *PLoS Biology* 11(8): e1001643.
doi: 10.1371/journal.pbio.1001643

Vinckenbosch N, Dupanloup I, Kaessmann H (2006). Evolutionary fate of retroposed gene copies in the human genome. *Proceedings of the National Academy of Sciences of the United States of America* 103(9): 3220-3225.
doi: 10.1073/pnas.0511307103

Vogel C, Chothia C (2006). Protein family expansions and biological complexity. *PLoS Computational Biology* 2(5): e48.
doi: 10.1371/journal.pcbi.0020048

Wade MJ, Brandvain Y (2009). Reversing mother's curse: Selection on male mitochondrial fitness effects. *Evolution* 63(4): 1084-1089.
doi: 10.1111/j.1558-5646.2009.00614.x

Wade MJ, Goodnight CJ (2006). Cyto-nuclear epistasis: Two-locus random genetic drift in hermaphroditic and dioecious species. *Evolution* 60(4): 643-659.
doi: -

Wang Y, et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research* 40(7): e49-e49.
doi: 10.1093/nar/gkr1293

Wang Z, Gerstein M, Snyder M (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics* 10(1): 57-63.
doi: 10.1038/nrg2484

Wang Z, et al. (2014). Temporal genomic evolution of bird sex chromosomes. *BMC Evolutionary Biology* 14(1): 1-12.
doi: 10.1186/s12862-014-0250-8

Werren JH (2011). Selfish genetic elements, genetic conflict, and evolutionary innovation. *Proceedings of the National Academy of Sciences of the United States of America* 108(Supplement 2): 10863-10870.
doi: 10.1073/pnas.1102343108

Wheeler TJ, Eddy SR (2013). nhmmer: DNA homology search with profile HMMs. *Bioinformatics* 29(19): 2487-2489.
doi: 10.1093/bioinformatics/btt403

White MA, Kitano J, Peichel CL (2015). Purifying selection maintains dosage-sensitive genes during degeneration of the threespine stickleback Y chromosome. *Molecular Biology and Evolution* 32(8): 1981-1995.
doi: 10.1093/molbev/msv078

Wilson G, et al. (2014). Best practices for scientific computing. *PLoS Biology* 12(1): e1001745.
doi: 10.1371/journal.pbio.1001745

Wolf JBW, Bryk J (2011). General lack of global dosage compensation in ZZ/ZW systems? Broadening the perspective with RNA-seq. *BMC Genomics* 12(1): 1-10.
doi: 10.1186/1471-2164-12-91

Wolfe K (2000). Robustness - it's not where you think it is. *Nature Genetics* 25(1): 3-4.
doi: 10.1038/75560

Wolfe KH (2001). Yesterday's polyploids and the mystery of diploidization. *Nature Reviews Genetics* 2(5): 333-341.
doi: 10.1038/35072009

Wolfe KH, Shields DC (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387(6634): 708-712.
doi: 10.1038/42711

Wolff JN, Nafisinia M, Sutovsky P, Ballard JWO (2013). Paternal transmission of mitochondrial DNA as an integral part of mitochondrial inheritance in metapopulations of *Drosophila simulans*. *Heredity* 110(1): 57-62.
doi: 10.1038/hdy.2012.60

Wright AE, et al. (2015a). Variation in promiscuity and sexual selection drives avian rate of Faster-Z evolution. *Molecular Ecology* 24(6): 1218-1235.
doi: 10.1111/mec.13113

Wright AE, Moghadam HK, Mank JE (2012). Trade-off between selection for dosage compensation and masculinization on the avian Z chromosome. *Genetics* 192(4): 1433-1445.
doi: 10.1534/genetics.112.145102

Wright AE, Zimmer F, Harrison PW, Mank JE (2015b). Conservation of regional variation in sex-specific sex chromosome regulation. *Genetics* 201(2): 587-598. **doi**: 10.1534/genetics.115.179234

Xie Y, et al. (2014). SOAPdenovo-Trans: *De novo* transcriptome assembly with short RNA-Seq reads. *Bioinformatics* 30(12): 1660-1666. **doi**: 10.1093/bioinformatics/btu077

Xiong Y, et al. (2010). RNA sequencing shows no dosage compensation of the active X-chromosome. *Nature Genetics* 42(12): 1043-1047. **doi**: 10.1038/ng.711

Yandell M, Ence D (2012). A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics* 13(5): 329-342. **doi**: 10.1038/nrg3174

Yang Y, Smith SA (2013). Optimizing *de novo* assembly of short-read RNA-seq data for phylogenomics. *BMC Genomics* 14(1): 1-11. **doi**: 10.1186/1471-2164-14-328

Yang Z (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution* 24(8): 1586-1591. **doi**: 10.1093/molbev/msm088

Yates A, et al. (2015). The Ensembl REST API: Ensembl data for any language. *Bioinformatics* 31(1): 143-145. **doi**: 10.1093/bioinformatics/btu613

Yee WKW, Sutton KL, Dowling DK (2013). In vivo male fertility is affected by naturally occurring mitochondrial haplotypes. *Current Biology* 23(2): R55-R56. **doi**: 10.1016/j.cub.2012.12.002

Zarrei M, MacDonald JR, Merico D, Scherer SW (2015). A copy number variation map of the human genome. *Nature Reviews Genetics* 16(3): 172-183. **doi**: 10.1038/nrg3871

Zhang H, Guillaume F, Engelstädter J (2012). The dynamics of mitochondrial mutations causing male infertility in spatially structured populations. *Evolution* 66(10): 3179-3188. **doi**: 10.1111/j.1558-5646.2012.01675.x

Zhang J (2003). Evolution by gene duplication: An update. *Trends in Ecology & Evolution* 18(6): 292-298. **doi**: 10.1016/S0169-5347(03)00033-8

Zhang J (2006). Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys. *Nature Genetics* 38(7): 819-823. **doi**: 10.1038/ng1812

Zhang J, Zhang Y-p, Rosenberg HF (2002). Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. *Nature Genetics* 30(4): 411-415. **doi**: 10.1038/ng852

Zhao S, Zhang B (2015). A comprehensive evaluation of Ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. *BMC Genomics* 16(1): 1-14.
doi: 10.1186/s12864-015-1308-8

Zhou F, et al. (2013). NBPF is a potential DNA-binding transcription factor that is directly regulated by NF- κ B. *The international Journal of Biochemistry & Cell Biology* 45(11): 2479-2490.
doi: 10.1016/j.biocel.2013.07.022

Zhou Q, et al. (2008). On the origin of new genes in *Drosophila*. *Genome Research* 18(9): 1446-1455.
doi: 10.1101/gr.076588.108

Zilles K, Rehkämper G. (1988). The brain, with special reference to the telencephalon. In. Orang-utan biology. p. 157-176.

Appendix

The following article was first published in

Genome Biology and Evolution

Zimmer F, Harrison PW, Dessimoz C, Mank JE (2016) Compensation of dosage-sensitive genes on the chicken Z chromosome. *Genome Biology and Evolution* 8 (4): 1233-1242, **doi**: 10.1093/gbe/evw075

Compensation of Dosage-Sensitive Genes on the Chicken Z Chromosome

Fabian Zimmer*[,1], Peter W. Harrison[1], Christophe Dessimoz[1,2,3], and Judith E. Mank[1]

[1]Department of Genetics Evolution and Environment, University College London, London, United Kingdom

[2]Department of Ecology and Evolution & Center for Integrative Genomics, University of Lausanne, Biophore 1015, Lausanne, Switzerland

[3]Swiss Institute of Bioinformatics, Biophore, 1015 Lausanne, Switzerland

*Corresponding author: E-mail: fabian.zimmer.12@ucl.ac.uk.

Accepted: March 26, 2016

Data deposition: This project has been deposited at the NCBI Short Read Archive under the accession SRP065394.

Abstract

In many diploid species, sex determination is linked to a pair of sex chromosomes that evolved from a pair of autosomes. In these organisms, the degeneration of the sex-limited Y or W chromosome causes a reduction in gene dose in the heterogametic sex for X- or Z-linked genes. Variations in gene dose are detrimental for large chromosomal regions when they span dosage-sensitive genes, and many organisms were thought to evolve complete mechanisms of dosage compensation to mitigate this. However, the recent realization that a wide variety of organisms lack complete mechanisms of sex chromosome dosage compensation has presented a perplexing question: How do organisms with incomplete dosage compensation avoid deleterious effects of gene dose differences between the sexes? Here we use expression data from the chicken (*Gallus gallus*) to show that ohnologs, duplicated genes known to be dosage-sensitive, are preferentially dosage-compensated on the chicken Z chromosome. Our results indicate that even in the absence of a complete and chromosome wide dosage compensation mechanism, dosage-sensitive genes are effectively dosage compensated on the Z chromosome.

Key words: dosage sensitivity, whole genome duplication, sex chromosomes, ohnologs.

Introduction

Heteromorphic sex chromosomes have evolved independently in many species (Bachtrog et al. 2014; Beukeboom and Perrin 2014). In some cases, recombination has been suppressed along the majority of the length of the sex chromosomes, leading to a large-scale loss of active genes from the sex-limited Y and W chromosomes (Charlesworth et al. 2005; Bachtrog et al. 2011). This results in large differences in size, with one large, gene-rich chromosome (X or Z chromosome), and one smaller chromosome, lacking many genes (Y or W chromosome).

The decay of Y and W chromosome gene content leads to differences in gene dose between the sexes, where the heterogametic sex has one half of the dose of all genes lost from the sex-limited chromosome compared with the homogametic sex. For many loci, gene dose correlates with gene expression (Pollack et al. 2002; Birchler et al. 2005; Torres et al. 2007; Malone et al. 2012), therefore the reduced gene dose on the X or Z chromosome should result in reduced gene expression in the heterogametic sex. When dosage-sensitive genes are affected, this could lead to a reduction in fitness in the heterogametic sex, and result in selective pressures favoring the evolution of dosage compensation mechanisms (Ohno 1967; Charlesworth 1978, 1996, 1998). These mechanisms should equalize the expression between the sex chromosomes and the autosomes, thereby restoring them to the ancestral level before the decay of gene content on the W or Y chromosome. Second, they should equalize the expression of individual dosage-sensitive genes between males and females.

Although it was once assumed that complete and global dosage compensation would always be associated with sex chromosome evolution (Ohno 1967), there is considerable variation in the mechanism and completeness of dosage compensation across species. For example, in *Drosophila melanogaster* (Conrad and Akhtar 2012) and *Caenorhabditis elegans* (Meyer 2010), dosage balance is achieved through regulatory mechanisms affecting the entire X chromosome (Straub and Becker 2007). In these cases, differences in gene dose of the sex chromosome are compensated for and expression is on average balanced between the sexes for the X chromosome, and between the single X and the diploid autosomes in males, the heterogametic sex. However, it is now clear that complete mechanisms of dosage compensation are rare, and many organisms, including birds (Ellegren et al. 2007; Itoh et al. 2007; Naurin et al. 2011; Wolf and Bryk 2011; Uebbing et al. 2013; Wright et al. 2015), snakes (Vicoso et al. 2013), many insects (Vicoso and Bachtrog 2015), and fish (Leder et al. 2010; Chen et al. 2014), have incomplete dosage compensation (reviewed in Mank 2013).

Incomplete dosage compensation was first documented in chicken (Ellegren et al. 2007; Itoh et al. 2007) and subsequently confirmed in several other avian species (Naurin et al. 2011; Wolf and Bryk 2011; Uebbing et al. 2013; Wright et al. 2015). In birds, which are a model for studies of incomplete dosage compensation, there is a significant reduction in average expression of the Z chromosomes in females, the heterogametic sex, relative to the autosomes as well as to the male Z chromosome average (Ellegren et al. 2007; Itoh et al. 2007; Wolf and Bryk 2011; Uebbing et al. 2013, 2015). The realization that many organisms with heteromorphic sex chromosomes have not in fact evolved complete and global dosage compensation mechanisms is perplexing as it is unclear how these organisms cope with negative dose effects. A reduction in gene dose often does not produce an observable difference in expression for many genes (Malone et al. 2012), and it was unclear whether certain loci are actively dosage-compensated or simply lack dose effects.

One possible explanation proposed by Mank and Ellegren (2008) is that instead of requiring a global mechanism of dosage compensation, the regulation of gene dose might occur on a gene-by-gene basis. A more targeted, local mechanism of dosage compensation should primarily affect the expression of dosage-sensitive genes (Mank et al. 2011). The role of dosage-sensitivity for the evolution of dosage compensation mechanisms has been discussed by a number of reviews (Mank 2013; Pessia et al. 2013; Ercan 2015; Veitia et al. 2015) and was investigated in a range of species. For example, in mammals X chromosomal expression is reduced compared with the autosomes in both males and females (Xiong et al. 2010; Julien et al. 2012), possibly as a consequence of X chromosome inactivation. However, dosage-sensitive genes, such as protein–complexes, show evidence of a higher degree of dosage-compensation (Lin et al. 2012; Pessia et al. 2012), compared with other gene categories. Recent studies in nematodes (Albritton et al. 2014) and fish (White et al. 2015) also showed similar patterns of compensated dosage-sensitive genes.

Dosage-sensitivity can result from interactions with other genes or gene products (Veitia 2004), such as in the case of transcription factors and large protein complexes (Papp et al. 2003). Individual duplications of these dosage-sensitive genes are likely to be rare, as they disrupt the stoichiometric balance and may disturb gene networks (Birchler et al. 2001; Papp et al. 2003; Birchler and Veitia 2012). However, dosage-sensitive genes should be preferentially retained after whole genome duplications (WGDs) (Edger and Pires 2009; Birchler and Veitia 2012). In contrast, dosage-insensitive genes that do not exhibit neo- or sub-functionalization are often lost after WGD (Dehal and Boore 2005). WGDs have occurred in a wide range of lineages (Wolfe and Shields 1997; Kellis et al. 2004; Dehal and Boore 2005; Cui et al. 2006; Van de Peer et al. 2009), including two rounds of WGD events roughly 500 MYA ago (Dehal and Boore 2005), which gave rise to roughly 16–34% of the chicken genome (Singh et al. 2015).

Preferentially retained gene duplicates originating from WGDs, also known as ohnologs (Wolfe 2000, 2001), are skewed toward gene families associated with dosage-sensitive functions such as signaling and development (Blomme et al. 2006) and protein–complexes (Makino et al. 2009). The dosage sensitivity of ohnologs (Blomme et al. 2006; Makino et al. 2009) is well established and makes them particularly useful in assessing the effectiveness of incomplete dosage compensation. We therefore use ohnologs to investigate the effectiveness of compensation on the chicken Z chromosome and to understand the evolution of incomplete sex chromosome dosage compensation mechanisms in general.

Results

We generated RNA-Seq gene expression profiles from multiple male and female biological replicates for four different tissues (spleen, heart, liver, and gonad) in chicken (*Gallus gallus*), recovering on average 17 million paired-end mappable reads per sample. We removed genes that were not expressed on average in all male and female above at least two counts per million (CPM). The number of genes expressed on the autosomes and Z chromosome for each tissue are shown in supplementary table S1, Supplementary Material online.

Incomplete Dosage Compensation in Females and Reduced Z Expression in Males

Dosage compensation has been assessed in a variety of ways, often depending on the system being studied. We used two approaches to assess dosage compensation status. First, complete dosage compensation should equalize female Z-linked and autosomal expression. Second, dosage compensation can also act on a local gene-by-gene basis, balancing the individual gene expression in males and females, which may be the dominant mechanism for dosage-sensitive genes.

Consistent with previous studies showing the incomplete dosage compensation in chicken, we detected lower average expression of Z-linked genes in comparison to autosomal genes in all female tissues (spleen $P < 0.0001$, Z-score = 11.19; heart $P < 0.0001$, Z-score = 11.22; liver $P < 0.0001$, Z-score = 8.88; ovaries $P < 0.0001$, Z-score = 9.20; Wilcoxon

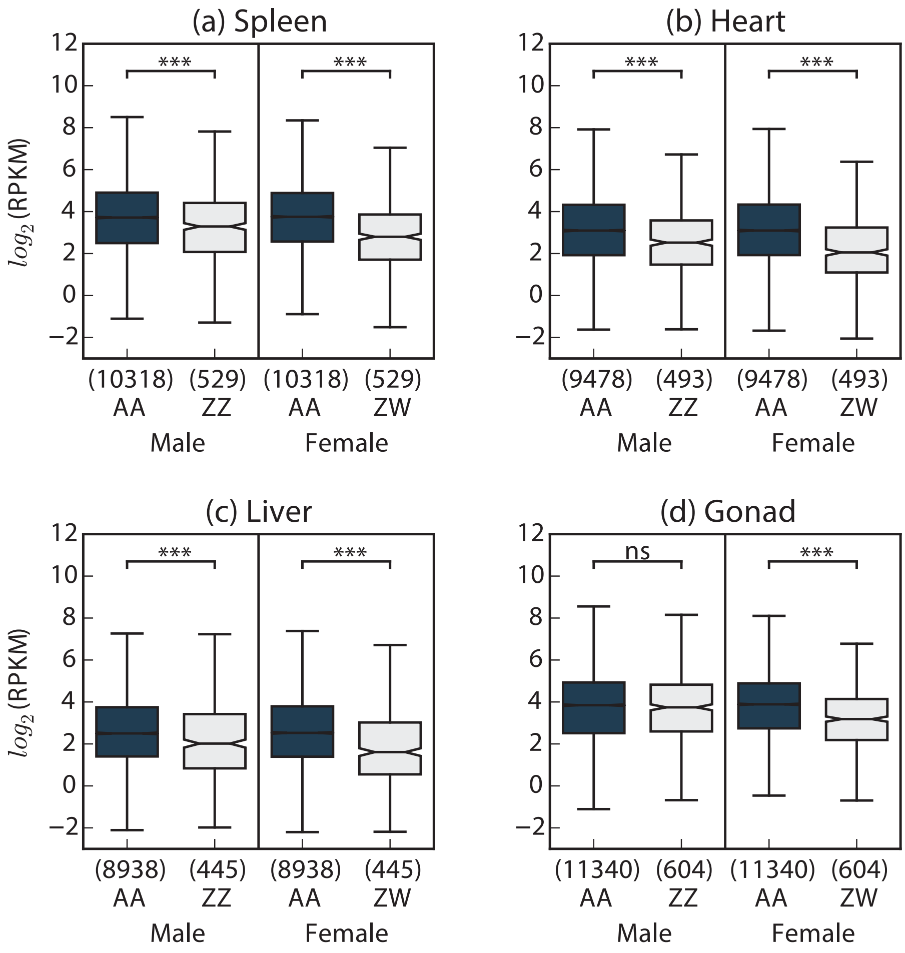


Fig. 1.—Comparison of gene expression measured for autosomal genes (dark grey) and Z-linked genes (light grey) in (*a*) spleen, (*b*) heart, (*c*) liver, and (*d*) gonad tissue in males and females. In all tissues, gene expression for Z-linked genes is significantly lower in comparison to autosomal genes in females. In males, gene expression of Z-linked genes is significantly lower in comparison to autosomal genes in all somatic tissues but not in gonad. Significance levels are indicated as stars (*$P < 0.05$, **$P < 0.001$, ***$P < 0.0001$), differences between distributions were tested using Wilcoxon Rank Sum tests. The number of genes expressed on the autosomes and Z chromosome(s) are given in brackets for each distribution. Boxes show the interquartile range, notches represent the median of the distribution and whiskers extend to 1.5 times the interquartile range (Q3 + 1.5 × IQR, Q1–1.5 × IQR). Outliers are not shown for clarity, but included in all statistical comparisons.

Rank Sum Test, fig. 1, supplementary fig. S1 and table S2, Supplementary Material online). We also expect that the average expression of the Z chromosomes in males is similar to the autosomal average, as two Z chromosomes are present. In line with this prediction, we find that the distribution of male expression is not significantly different to the autosomes in testes ($P = 0.79$, *Z*-score = 0.27, Wilcoxon Rank Sum Test). However, a previous study has indicated that in some tissues, expression of the Z in males is also less than the autosomal average (Julien et al. 2012), and we also recovered a significant reduction in average expression of Z-linked loci compared with average autosomal expression in all somatic tissues in males (spleen $P < 0.0001$, *Z*-score = 5.50; heart $P < 0.0001$, *Z*-score = 6.69; liver $P < 0.0001$, *Z*-score = 5.02; Wilcoxon Rank Sum Test). When we compared the average expression level of all autosomes and the Z chromosomes, it is clear that the Z chromosome expression in both males and females is outside the autosomal spectrum for all somatic tissues (supplementary fig. S1, Supplementary Material online).

One possible explanation for the low *Z* expression could be the inclusion of lowly expressed genes, but the median *Z:A* ratios for males (*ZZ:AA*) and females (*Z:AA*) across a range of higher CPM expression thresholds (supplementary fig. S2, Supplementary Material online) is similar, suggesting that a minimum CPM threshold > 2 is effective in filtering out lowly expressed genes. The difference in male and female Z-linked gene expression is also robust across expression quartiles, except for gonad expression quartile one (supplementary fig. S3, Supplementary Material online). The reduction in Z expression in males is also consistent with the possible inactivation of one Z chromosome in males, analogous to the X inactivation observed in therian females (Cooper et al. 1993; Deakin et al. 2009). Male Z chromosome inactivation has been suggested by previous work on a limited number of Z-linked loci (Livernois et al. 2013) and we investigated the potential for Z inactivation using our RNA-Seq data. If one copy of the Z chromosome were partially inactivated in males, we would expect to find SNPs with a significantly greater contribution to the total expression from one allele at heterozygous sites. Our analyses of allele-specific expression (ASE) indicate that only a limited number of Z-linked genes exhibit ASE, and there is no robust evidence that the proportion is greater than that observed for the autosomes (Supplementary Material online). This suggests that the reduction in male expression on the Z chromosome is not due to chromosomal inactivation.

Ohnologs Are Preferentially Dosage-Compensated

If incomplete dosage compensation is sufficient for compensating dosage-sensitive genes, we might expect the proportion of dosage-compensated ohnologs on the Z chromosome to be higher in comparison to nonohnologs. We tested whether ohnologs are more often dosage-compensated using our expression data and ohnologs obtained from the OhnologsDB (Singh et al. 2015). The chicken genome contains 5,228 (33.71%) annotated ohnologs, of which 223 are annotated on the Z chromosome. Z chromosome ohnologs show over-enrichment for Gene Ontology terms compared with all genes, such as cell motility and locomotion, which may be important in dosage sensitivity (supplementary table S3, Supplementary Material online).

In order to determine whether ohnologs are preferentially dosage-compensated, we first compared the $\log_2$ fold change between female and male expressions for Z-linked ohnologs and nonohnologs (fig. 2). The difference in expression between females and males ($\log_2$FC) was significantly lower for ohnologs than nonohnologs (spleen $P < 0.0001$, Z-score = 5.95; heart $P < 0.0001$, Z-score = 4.57; liver $P < 0.0001$, Z-score = 5.22; gonad $P < 0.0001$, Z-score = 4.89; Wilcoxon Rank Sum Test), suggesting a higher degree of dosage compensation. In addition, the proportion of dosage-compensated ohnologs ($\log_2$FC range from −0.5 to 0.5) was significantly higher when compared with nonohnologs in all tissues (P-value < 0.0001 in all comparisons; Fisher's Exact test, table 1). This is also the case when we used a wider range of $\log_2$FC (−0.6 to 0.6), similar to the mean expression change for female one-dose genes reported by Malone et al. (2012) (supplementary table S4, Supplementary Material online). In addition, we used the strict set of ohnologs from the OhnologsDB, with 2,489 ohnologs annotated in the chicken genome and 106 on the Z chromosome, recovering similar results (supplementary fig. S4 and Table S5, Supplementary Material online).

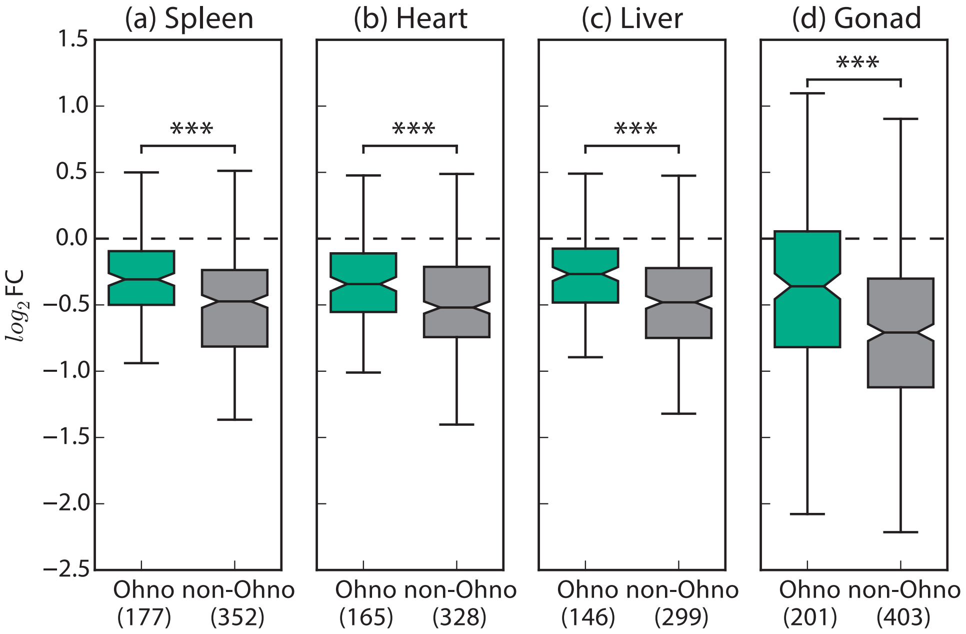


FIG. 2.—Comparison of $\log_2$-transformed fold change between female and male expressions for ohnologs (green) and nonohnologs (grey) on the Z chromosome in (*a*) spleen, (*b*) heart, (*c*) liver, and (*d*) gonad. The number of genes in the distributions is given in brackets. Negative fold changes indicate higher male expression; positive fold changes indicate stronger female expression. Significance levels are indicated as stars (*$P < 0.05$, **$P < 0.001$, ***$P < 0.0001$), differences between distributions were tested using Wilcoxon Rank Sum tests. Outliers are not shown for clarity, but included in all statistical comparisons.

Table 1
Contingency Tables for All Four Tissues, Comparing the Proportion of Dosage-Compensated (DC) and Uncompensated (U) Ohnologs to Non-ohnologs Using a Fisher's Exact Test

	Ohnolog		Non-ohnolog		*P* value	Odds ratio
	DC	U	DC	U		
Spleen	126 (71.19%)	51 (28.81%)	180 (51.14%)	172 (48.86%)	**1.08×10^{-5}**	2.36
Heart	111 (67.27%)	54 (32.73%)	152 (46.34%)	176 (53.66%)	**1.06×10^{-5}**	2.38
Liver	105 (71.92%)	41 (28.08%)	147 (49.16%)	152 (50.84%)	**6.52×10^{-6}**	2.65
Gonad	86 (42.79%)	115 (57.21%)	103 (25.56%)	300 (74.44%)	**2.57×10^{-5}**	2.18

NOTE—Significant *P* values are reported in bold.

An alternative explanation for the high degree of dosage compensation among ohnologs is that all paralogs, even those that originate in single-gene duplications, are dosage-compensated. We tested this hypothesis by extracting Z-linked paralogs from the Ensembl database (Cunningham et al. 2015) that originated in single-gene duplication events. These paralogs do not show a higher proportion of dosage compensation ($P > 0.05$ in all comparisons; Fisher's Exact test; supplementary table S6, Supplementary Material online) compared with all other genes on the Z chromosome. This indicates that the higher degree of dosage compensation among ohnologs is not a property of paralogs in general, and that the mode of duplication has an important impact on the evolution of gene-by-gene dosage compensation.

Older Z Chromosome Parts Contain Fewer Ohnologs

Sex chromosome divergence can drive the movement of some gene classes off the sex chromosomes (Emerson et al. 2004; Potrzebowski et al. 2008; Vibranovski et al. 2009) and we might expect an out of Z migration for dosage-sensitive genes. Overall, the proportion of ohnologs is not significantly different between the Z (764 coding genes) and the genomic background (14,744 coding genes) ($P = 0.19$, odds ratio = 0.89; Fisher's Exact test), suggesting that the Z chromosome is not depleted of ohnologs and that dosage-sensitive gene have not moved off the Z. However, the Z chromosome contains at least four strata, where recombination was suppressed between the Z and W at different times, spanning roughly 130 million years (Wright et al. 2012). We divided the chromosome into an old and young parts along the border of stratum 3, resulting in two almost equally sized regions of the Z chromosome. Given 223 ohnologs located on the Z chromosome, we expect that half of these would be located in the old and half in the young part of the chromosome. However, the number of ohnologs in the older half of the chromosome is significantly less than expected ($\chi^2 = 22.605$, $P < 0.0001$; Chi-square test), and also significantly less when accounting for the difference in gene content ($P < 0.05$, odds ratio = 0.62; Fisher's Exact test). This could indicate that some ohnologs may have relocated during the early evolution of the Z chromosome. When we compared the proportion of dosage-compensated ohnologs between old and young parts of the Z chromosome, we do not detect a significantly higher proportion of dosage-compensated ohnologs in older parts ($P > 0.05$ in all comparisons; Fisher's Exact test), suggesting that dosage compensation of ohnologs occurs relatively quickly following W chromosome gene loss. Alternatively, this bias could be an artifact of the ancestral ohnolog distribution, as the WGD events precede the formation of the sex chromosome system.

Dosage Compensation of Ohnologs across Tissues

The degree of dosage compensation is similar in all somatic tissues ($P > 0.05$ in all comparisons; Fisher's Exact test; supplementary table S7, Supplementary Material online), and greater in the soma compared with the gonad ($P < 0.0001$ in all comparisons; Fisher's Exact test; supplementary table S7, Supplementary Material online). Tissues can be seen as a form of functional compartmentalization, and the same gene can show a diverse range of expression patterns in different tissues. For this reason, similar overall dosage compensation could hide an underlying pattern of pleiotropic expression. Dosage sensitivity may in fact be tissue dependent and can result in gene-by-gene dosage compensation (Mank and Ellegren 2008).

We also investigated the overlap of dosage-compensated ohnologs across tissues. A set of 68 of 223 ohnologs was dosage-compensated in all somatic tissues; however, we detected substantial variation (fig. 3). Of the 68 ohnologs that are dosage-compensated in all somatic tissues, only 36 are also dosage-compensated in gonad, showing that only a small core set of ohnologs are dosage-sensitive across all tissues. In gonad, a unique set of 50 ohnologs was dosage-compensated. In combination with the overall lower degree of dosage compensation in gonad, this suggests different dosage compensation patterns when compared with the somatic tissues.

Discussion

Our analyses of dosage compensation and ohnologs on the chicken Z chromosome provide novel insights into the nature of incomplete dosage compensation. We confirm previous

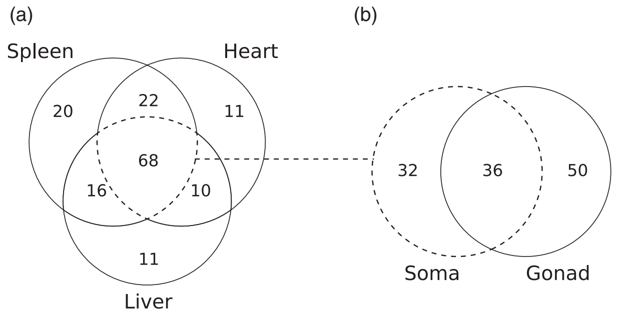


Fig. 3.—(*a*) Overlap between dosage-compensated ohnologs in the three somatic tissues. (*b*) Overlap between dosage-compensated genes in the soma (spleen, heart, and liver) and gonad tissue. Circles represent the total of dosage-compensated ohnologs in a tissue and numbers indicate the overlap between sets.

reports of incomplete dosage compensation in chicken (Ellegren et al. 2007; Itoh et al. 2007; Uebbing et al. 2015) and show that ohnologs are preferentially dosage-compensated on the chicken Z chromosome, indicating that incomplete dosage compensation can effectively balance dosage-sensitive genes. Even though the average expression of the Z chromosome is consistently lower in females as a function of incomplete dosage compensation, a considerable number of Z-linked genes show equal expression between males and females. Moreover, selection for compensation of dosage-sensitive genes appears to act relatively quickly, as there is no significant difference in the proportion of dosage-compensated ohnologs in younger regions of the avian Z chromosome compared with older regions.

The X chromosomal expression in mammals is reduced compared with the autosomes, potentially as a consequence of X inactivation (Xiong et al. 2010; Julien et al. 2012). It has been suggested that selection for the compensation of dosage-sensitive genes could have driven the evolution of X inactivation in therian mammals. Similarly, we also observe a reduction in Z expression in somatic tissues in males (Itoh et al. 2007). The reduced expression of the Z chromosome compared with the autosomes in males is not as pronounced as in females (fig. 1, supplementary table S2, Supplementary Material online) and there are several possible explanations for this pattern. The reduction has been suggested to result by partial *Z* inactivation that affects parts of the chromosome (Livernois et al. 2013; Graves 2014). However, our assessment of ASE suggests that inactivation is not a major mechanism affecting Z chromosome expression in males. An alternative explanation for the lower Z expression may be that the ancestral expression level of the Z chromosome, before the differentiation of the sex chromosomes, was already on average on the lower end of the expression spectrum (Brawand et al. 2011; Julien et al. 2012). Finally, it is possible that dosage sensitive genes have moved off the Z, as the mammalian X chromosome is depleted of genes requiring high transcription rates as a result of haploid expression in females (Hurst et al. 2015). Our analysis suggests that although there is some potential for movement of dosage-sensitive genes off the Z chromosome, the effect is confined to the oldest regions of the Z chromosome and is not substantial enough to explain the reduced expression in males.

It is important to keep in mind that the detection of ohnologs in vertebrate genomes remains challenging due to the age of the two rounds of WGD. All tools for the detection of ohnologs depend on the analysis of preserved gene order (synteny) among paralogs to distinguish single-gene duplicates from WGD. Large intra-genomic rearrangements may complicate these analyses, and may result in the underestimation of the number of ohnologs. Avian genomes, however, are relatively stable and compact, with fewer repeats and more coding DNA compared with other amniotes (Hillier et al. 2004; Ellegren 2005; Organ et al. 2007), suggesting that these issues are less prevalent. In addition, the detection of ohnologs depends on the selection of one or more outgroups that did not undergo a WGD to distinguish between genes that were duplicated before the WGD events. The outgroup selection can influence the number of ohnologs (Makino and McLysaght 2010) and the OhnologsDB mitigates that issue by using multiple outgroups.

Conclusion

Our results are consistent with gene-by-gene dosage compensation (Mank and Ellegren 2008; Mank 2013; Uebbing et al. 2013) and demonstrate that selection for dosage compensation of ohnologs does not necessitate the evolution of a global dosage compensation mechanism. This in turn leads to the interesting question why some organisms exhibit complex mechanisms of complete dosage compensation that require regulation of the entire X chromosome when such mechanisms are not necessarily evolutionarily required.

Methods

RNA-Seq Analysis and Gene Expression Estimates

We collected heart, liver, and spleen samples from White Leghorn chicken (*G. gallus*) embryonic day 19 eggs incubated under standard conditions. Embryos were sexed visually and based on expression of W-linked genes. For each tissue, four biological samples were collected for both males and females. One female liver sample was excluded from the analyses because it showed only spurious W expression and when investigating the *Z:A* ratio it was clearly masculinized. All samples were first stored in RNAlater (Qiagen) and then total RNA was extracted (Qiagen Animal Tissue RNA kit).

Library construction and Illumina sequencing was done at the Wellcome Trust Centre of Human Genetics (WTCHG), Oxford. Each sample was normalized to 2.5 µg total RNA prior to a PolyA isolation using an NEB Magnetic mRNA Isolation Kit. PCR was carried out over 15 cycles using custom-indexed primers (WTCHG). Libraries were quality controlled with picogreen and tapestation, and were subsequently normalized equimolarly into 12-plex pools for Illumina HiSeq sequencing. Heart, liver, and spleen samples were sequenced using an Illumina HiSeq 2000 as paired-end 100-bp reads. 51-bp paired end reads of gonadal samples from the same development stage were obtained from Moghadam et al. (2012).

We trimmed each library using Trimmomatic v0.22 (Lohse et al. 2012) removing leading and trailing bases with a Phred score <3 and trimming using a sliding window approach when the average Phred score over four bases was <15. Reads were kept if they were at least 36 bases after trimming. Libraries were quality-inspected manually using FASTQC v0.10.1 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The trimmed libraries were aligned against the chicken reference genome Ensembl version 75 Galgal4 (Cunningham et al. 2015) using TopHat v2.0.11 (Kim et al. 2013) and bowtie2 v2.2.2 (Langmead and Salzberg 2012) allowing five mismatches to the reference genome, with on average 17 million paired-end mappable reads per sample. Multi-mapping reads were removed and we then sorted and indexed the resulting alignment files for each library separately using Samtools v0.1.18/9 (Li et al. 2009).

We extracted reads mapping to annotated genes using HTseq-Count v0.6.1p1 (Anders et al. 2014) and normalized all tissues separately using the trimmed mean of M-values method available in edgeR v3.2.4 (Robinson et al. 2010). We estimated differential expression between males and females in all tissues using edgeR's exactTest method and exported the $\log_2$ fold change ($\log_2$FC; female–male expression), average $\log_2$ count per million (logCPM), FDR corrected *P*-values from the exactTest function and individual CPM values for all samples and genes. Genes were only included when the average CPM was >2 across all males and females, filtering out loci with low expression. When comparing groups of genes to each other, we normalized the CPM values by gene length, resulting in reads per kilobase of transcript per million mapped reads values (RPKM). Only genes annotated to the autosomes and the Z chromosome were assessed. Individual genes were defined as dosage-compensated on the Z chromosome if the female:male $\log_2$ fold change ranged from -0.5 to 0.5 (Wright et al. 2015). We defined genes as sex-biased if the edgeR exactTest was significant after FDR correction ($q < 0.05$) and the $\log_2$ fold change was >1 for female-biased genes or <-1 for male-biased genes.

Identification of Ohnologs and Other Paralogs

We used the Ohnologs database (http://ohnologs.curie.fr/) (Singh et al. 2015) to obtain ohnologs present in the chicken genome. We used the relaxed set of ohnologs as the primary dataset, in order to maximize the number of ohnologs. In addition, we used the Ensembl REST API (accessed February 2015) (Yates et al. 2015) to identify all paralogs in the chicken genome, which also includes those homologs originated in single-gene duplications.

Functional Annotation of Ohnologs

We used the G:profiler toolkit (Reimand et al. 2011) to perform GO Term (Ashburner et al. 2000) overrepresentation analyses. All ohnologs on the Z chromosome were provided as an input list and compared with the entire genomic background, using only genes with annotated GO terms in the comparison. Standard settings were used and GO Terms were only considered if they had a significant *P* value after multiple testing correction via G:Profiler's G:SCS method (*P* value <0.05). We additionally used the CORUM database (Ruepp et al. 2010), version from February 2012, to annotate protein complexes in the chicken genome. The CORUM database contains only mammalian data and we used the Ensembl REST API (Yates et al. 2015) to detect the corresponding chicken homologs, where possible.

SNP Calling and Estimation of ASE

In order to detect ASE from RNA-Seq data we modified a pipeline from Quinn et al. (2014). As we were interested in detecting ASE on the Z chromosome, we only called SNPs in the homogametic sex (males) for each tissue. SNPs were called using Samtools mpileup v0.1.18 (Li et al. 2009) and VarScan2 v2.3.6 (Koboldt et al. 2012). SNPs were called separately for each tissue using all four available male samples. We required minimum coverage of 2 and minimum Phred score of 20 (–min-avg-qual 20) to call an SNP and also required a minimum frequency of 0.9 to call a homozygote (–min-freq-for-hom 0.9). The resulting variant call formatted files were then filtered further to remove noise and increase SNP call confidence. In a first step, we filtered out SNPs using a combination of a fixed minimum threshold of 17 reads per site (the combination of major and minor allele) in all samples, as our power analysis indicates that a 17 read coverage for an SNP results in 73% power to detect allele specific-expression and also excluded all SNPs with more than two alleles. We additionally used a variable threshold that accounts for the likelihood of observing a second allele because of sequencing errors an error probability of 1 in 100 (Quinn et al. 2014) and a maximum coverage of 100,000. RNA-Seq data have an intrinsic bias for the estimation of ASE, because those reads that resemble the reference genome have a higher probability of aligning successfully. In order to remove this bias, we eliminated clusters of SNPs if there were >5 SNPs in a window of 100 bp (Stevenson et al. 2013). We used BEDtools intersect v2.20.1 (Quinlan and Hall 2010) to filter out all SNPs that were not located in a known transcript.

If both chromosomes are active to the same degree, we expect that the probability of observing reads from one or the other chromosome is 0.5. We therefore used a two-tailed binomial test to show significant deviations from this expected distribution ($P < 0.05$). Binomial tests were corrected for multiple testing on the autosomes, because of the larger number of testable sites. In order to account for the fact that binomial tests will be significant even for very small deviations in the observed distribution when the sample size, in our case the alignment depth, is big enough, we also employed a minimum threshold of 70% reads stemming from one allele to call significant ASE. In addition, we used a power analysis to ensure that our ability to detect ASE is sufficient. At a minimum coverage of 17 reads per site our power for detecting ASE is >73%, which suggests that we are able to detect patterns of ASE successfully in most cases. We only included genes in the analysis if at least one SNP showed consistent ASE across all samples.

All analyses and statistical comparisons were performed using Python, Matplotlib (Hunter 2007) and R (R Core Team 2015), code and iPython notebooks (Pérez and Granger 2007) are available on GitHub at https://github.com/qfma/ohnolog-dc. All sequencing data used in the analyses are available in the NCBI Short Read Archive under accession number SRP065394.

Supplementary Material

Supplementary tables S1–S7 and figures S1–S4 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

We thank Stephen Montgomery, Alison Wright, Natasha Bloch, Vicencio Oostra, Iulia Darolti, and Kevin Fowler for helpful comments on the manuscript. This study was supported by a European Research Council Framework 7 grant (agreement 260233) to J.E.M, University College London and a Swiss National Science Foundation grant to C.D. (PP00P3_150654). Sequencing was performed by the Wellcome Trust Centre for Human Genetics Sequencing Hub (funded by Wellcome Trust Grant reference 090532/Z09/Z and MRC Hub Grant G0900747 91070). The authors acknowledge the use of the UCL *Legion* High Performance Computing Facility (Legion@UCL) and associated support services, in the completion of this study.

Literature Cited

Albritton SE, et al. 2014. Sex-biased gene expression and evolution of the X chromosome in nematodes. Genetics doi: 10.1534/genetics.114.163311.

Anders S, Pyl PT, Huber W. 2014. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics. 31(2):166–169.

Ashburner M, et al. 2000. Gene Ontology: tool for the unification of biology. Nat Genet. 25:25–29.

Bachtrog D, et al. 2011. Are all sex chromosomes created equal? Trends Genet. 27:350–357.

Bachtrog D, et al. 2014. Sex determination: why so many ways of doing it? PLoS Biol. 12:e1001899.

Beukeboom L, Perrin N. 2014. The evolution of sex determination. Cambridge (MA): Oxford University Press.

Birchler JA, Bhadra U, Bhadra MP, Auger DL. 2001. Dosage-dependent gene regulation in multicellular eukaryotes: implications for dosage compensation, aneuploid syndromes, and quantitative traits. Dev Biol. 234:275–288.

Birchler JA, Riddle NC, Auger DL, Veitia RA. 2005. Dosage balance in gene regulation: biological implications. Trends Genet. 21: 219–226.

Birchler JA, Veitia RA. 2012. Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. Proc Natl Acad Sci U S A. 109:14746–14753.

Blomme T, et al. 2006. The gain and loss of genes during 600 million years of vertebrate evolution. Genome Biol. 7:R43.

Brawand D, et al. 2011. The evolution of gene expression levels in mammalian organs. Nature 478:343–348.

Charlesworth B. 1978. Model for evolution of Y chromosomes and dosage compensation. Proc Natl Acad Sci U S A. 75:5618–5622.

Charlesworth B. 1996. The evolution of chromosomal sex determination and dosage compensation. Curr Biol. 6:149–162.

Charlesworth B. 1998. Sex chromosomes: evolving dosage compensation. Curr Biol. 8:R931–R933.

Charlesworth D, Charlesworth B, Marais G. 2005. Steps in the evolution of heteromorphic sex chromosomes. Heredity 95:118–128.

Chen S, et al. 2014. Whole-genome sequence of a flatfish provides insights into ZW sex chromosome evolution and adaptation to a benthic lifestyle. Nat Genet. 46:253–260.

Conrad T, Akhtar A. 2012. Dosage compensation in *Drosophila melanogaster*: epigenetic fine-tuning of chromosome-wide transcription. Nat Rev Genet. 13:123–134.

Cooper DW, Johnston PG, Watson JM, Graves JAM. 1993. X-inactivation in marsupials and monotremes. Semin Dev Biol. 4:117–128.

Cui L, et al. 2006. Widespread genome duplications throughout the history of flowering plants. Genome Res. 16:738–749.

Cunningham F, et al. 2015. Ensembl 2015. Nucleic Acids Res. 43:D662–D669.

Deakin J, Chaumeil J, Hore T, Marshall Graves J. 2009. Unravelling the evolutionary origins of X chromosome inactivation in mammals: insights from marsupials and monotremes. Chromosome Res. 17:671–685.

Dehal P, Boore JL. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol. 3:e314.

Edger P, Pires JC. 2009. Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. Chromosome Res. 17:699–717.

Ellegren H. 2005. The avian genome uncovered. Trends Ecol Evol. 20:180–186.

Ellegren H, et al. 2007. Faced with inequality: chicken do not have a general dosage compensation of sex-linked genes. BMC Biol. 5:40.

Emerson JJ, Kaessmann H, Betrán E, Long M. 2004. Extensive gene traffic on the mammalian X chromosome. Science 303:537–540.

Ercan S. 2015. Mechanisms of X chromosome dosage compensation. J Genomics 3:1–19.

Graves JM. 2014. Avian sex, sex chromosomes, and dosage compensation in the age of genomics. Chromosome Res. 22:45–57.

Hillier LW, et al. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432:695–716.

Hunter JD. 2007. Matplotlib: a 2D graphics environment. Comput Sci Eng. 9:90–95.

Hurst L, Tashakkori Ghanbarian A, Forrest ARR, Huminiecki L. 2015. The constrained maximal expression level owing to haploidy shapes gene content on the mammalian X chromosome. PLoS Biol. 13(12):e1002315.

Itoh Y, et al. 2007. Dosage compensation is less effective in birds than in mammals. J Biol. 6:2.

Julien P, et al. 2012. Mechanisms and evolutionary patterns of mammalian and avian dosage compensation. PLoS Biol. 10:e1001328.

Kellis M, Birren BW, Lander ES. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae.* Nature 428:617–624.

Kim D, et al. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14:R36.

Koboldt DC, et al. 2012. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 22:568–576.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359.

Leder EH, et al. 2010. Female-biased expression on the X chromosome as a key step in sex chromosome evolution in threespine sticklebacks. Mol Biol Evol. 27:1495–1503.

Li H, et al. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079.

Lin F, Xing K, Zhang J, He X. 2012. Expression reduction in mammalian X chromosome evolution refutes Ohno's hypothesis of dosage compensation. Proc Natl Acad Sci U S A. 109:11752–11757.

Livernois AM, et al. 2013. Independent evolution of transcriptional inactivation on sex chromosomes in birds and mammals. PLoS Genet. 9:e1003635.

Lohse M, et al. 2012. RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. Nucleic Acids Res. 40:W622–W627.

Makino T, Hokamp K, McLysaght A. 2009. The complex relationship of gene duplication and essentiality. Trends Genet. 25:152–155.

Makino T, McLysaght A. 2010. Ohnologs in the human genome are dosage balanced and frequently associated with disease. Proc Natl Acad Sci U S A. 107:9270–9274.

Malone J, et al. 2012. Mediation of *Drosophila* autosomal dosage effects and compensation by network interactions. Genome Biol. 13:R28.

Mank JE. 2013. Sex chromosome dosage compensation: definitely not for everyone. Trends Genet. 29:677–683.

Mank JE, Ellegren H. 2008. All dosage compensation is local: gene-by-gene regulation of sex-biased expression on the chicken Z chromosome. Heredity 102:312–320.

Mank JE, Hosken DJ, Wedell N. 2011. Some inconvenient truths about sex chromosome dosage compensation and the potential role of sexual conflict. Evolution 65:2133–2144.

Meyer BJ. 2010. Targeting X chromosomes for repression. Curr Opin Genet Dev. 20:179–189.

Moghadam HK, et al. 2012. W chromosome expression responds to female-specific selection. Proc Natl Acad Sci U S A. 109:8207–8211.

Naurin S, et al. 2011. The sex-biased brain: sexual dimorphism in gene expression in two species of songbirds. BMC Genomics 12:37.

Ohno S. 1967. Sex chromosomes and sex-linked genes. Berlin: Springer.

Organ CL, et al. 2007. Origin of avian genome size and structure in non-avian dinosaurs. Nature 446:180–184.

Papp B, Pál C, Hurst LD. 2003. Dosage sensitivity and the evolution of gene families in yeast. Nature 424:194–197.

Pérez F, Granger BE. 2007. IPython: a system for interactive scientific computing. Comput Sci Eng. 9:21–29.

Pessia E, Engelstädter J, Marais GAB. 2013. The evolution of X chromosome inactivation in mammals: the demise of Ohno's hypothesis? Cell Mol Life Sci. 71(8):1383–1394.

Pessia E, et al. 2012. Mammalian X chromosome inactivation evolved as a dosage-compensation mechanism for dosage-sensitive genes on the X chromosome. Proc Natl Acad Sci U S A. 109:5346–5351.

Pollack JR, et al. 2002. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. Proc Natl Acad Sci U S A. 99:12963–12968.

Potrzebowski L, et al. 2008. Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. PLoS Biol. 6:e80.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842.

Quinn A, Juneja P, Jiggins FM. 2014. Estimates of allele-specific expression in *Drosophila* with a single genome sequence and RNA-seq data. Bioinformatics 30:2603–2610.

R Core Team 2015. R: a language and environment for statistical computing. R Foundation for Statistical Computing.

Reimand J, Arak T, Vilo J. 2011. g:Profiler—a web server for functional interpretation of gene lists (2011 update). Nucleic Acids Res. 39:W307–W315.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139–140.

Ruepp A, et al. 2010. CORUM: the comprehensive resource of mammalian protein complexes—2009. Nucleic Acids Res. 38:D497–D501.

Singh PP, Arora J, Isambert H. 2015. Identification of ohnolog genes originating from whole genome duplication in early vertebrates, based on synteny comparison across multiple genomes. PLoS Comput. Biol. 11:e1004394.

Stevenson K, Coolon J, Wittkopp P. 2013. Sources of bias in measures of allele-specific expression derived from RNA-seq data aligned to a single reference genome. BMC Genomics 14:536.

Straub T, Becker PB. 2007. Dosage compensation: the beginning and end of generalization. Nat Rev Genet. 8:47–57.

Torres EM, et al. 2007. Effects of aneuploidy on cellular physiology and cell division in haploid yeast. Science 317:916–924.

Uebbing S, et al. 2015. Quantitative mass spectrometry reveals partial translational regulation for dosage compensation in chicken. Mol Biol Evol. 32:2716–2725.

Uebbing S, Künstner A, Mäkinen H, Ellegren H. 2013. Transcriptome sequencing reveals the character of incomplete dosage compensation across multiple tissues in flycatchers. Genome Biol Evol. 5:1555–1566.

Van de Peer Y, Maere S, Meyer A. 2009. The evolutionary significance of ancient genome duplications. Nat Rev Genet. 10:725–732.

Veitia RA. 2004. Gene dosage balance in cellular pathways: implications for dominance and gene duplicability. Genetics 168:569–574.

Veitia RA, Veyrunes F, Bottani S, Birchler JA. 2015. X chromosome inactivation and active X upregulation in therian mammals: facts, questions, and hypotheses. J Mol Cell Biol. 7:2–11.

Vibranovski MD, Zhang Y, Long M. 2009. General gene movement off the X chromosome in the *Drosophila* genus. Genome Res. 19:897–903.

Vicoso B, et al. 2013. Comparative sex chromosome genomics in snakes: differentiation, evolutionary strata, and lack of global dosage compensation. PLoS Biol. 11:e1001643.

Vicoso B, Bachtrog D. 2015. Numerous transitions of sex chromosomes in diptera. PLoS Biol. 13:e1002078.

White MA, Kitano J, Peichel CL. 2015. Purifying selection maintains dosage-sensitive genes during degeneration of the threespine stickleback Y chromosome. Mol Biol Evol. 32:1981–1995.

Wolf J, Bryk J. 2011. General lack of global dosage compensation in ZZ/ZW systems? Broadening the perspective with RNA-seq. BMC Genomics 12:91.

Wolfe K. 2000. Robustness—it's not where you think it is. Nat Genet. 25:3–4.

Wolfe KH. 2001. Yesterday's polyploids and the mystery of diploidization. Nat Rev Genet. 2:333–341.

Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387:708–713.

Wright AE, Moghadam HK, Mank JE. 2012. Trade-off between selection for dosage compensation and masculinization on the avian Z chromosome. Genetics 192:1433–1445.

Wright AE, Zimmer F, Harrison PW, Mank JE. 2015. Conservation of regional variation in sex-specific sex chromosome regulation. Genetics 201:587–598.

Xiong Y, et al. 2010. RNA sequencing shows no dosage compensation of the active X-chromosome. Nat Genet. 42:1043–1047.

Yates A, et al. 2015. The ensembl REST API: ensembl data for any language. Bioinformatics 31:143–145.

Associate editor: Soojin Yi

The following article was first published in

Genome Biology and Evolution

Dean R, **Zimmer F**, Mank JE 2014**.** The potential role of sexual conflict and sexual selection in shaping the genomic distribution of mito-nuclear genes. *Genome Biology and Evolution* 6(5): 1096-1104, **doi**: 10.1093/gbe/evu063

The Potential Role of Sexual Conflict and Sexual Selection in Shaping the Genomic Distribution of Mito-nuclear Genes

Rebecca Dean, Fabian Zimmer, and Judith E. Mank*

Department of Genetics, Evolution, and Environment, University College London, London, United Kingdom

*Corresponding author: E-mail: judith.mank@ucl.ac.uk.

Accepted: March 12, 2014

Abstract

Mitochondrial interactions with the nuclear genome represent one of life's most important co-evolved mutualisms. In many organisms, mitochondria are maternally inherited, and in these cases, co-transmission between the mitochondrial and nuclear genes differs across different parts of the nuclear genome, with genes on the X chromosome having two-third probability of co-transmission, compared with one-half for genes on autosomes. These asymmetrical inheritance patterns of mitochondria and different parts of the nuclear genome have the potential to put certain gene combinations in inter-genomic co-adaptation or conflict. Previous work in mammals found strong evidence that the X chromosome has a dearth of genes that interact with the mitochondria (mito-nuclear genes), suggesting that inter-genomic conflict might drive genes off the X onto the autosomes for their male-beneficial effects. Here, we developed this idea to test coadaptation and conflict between mito-nuclear gene combinations across phylogenetically independent sex chromosomes on a far broader scale. We found that, in addition to therian mammals, only *Caenorhabditis elegans* showed an under-representation of mito-nuclear genes on the sex chromosomes. The remaining species studied showed no overall bias in their distribution of mito-nuclear genes. We discuss possible factors other than inter-genomic conflict that might drive the genomic distribution of mito-nuclear genes.

Key words: X chromosome, Z chromosome, sexual conflict, Haldane's sieve, OXPHOS.

Introduction

The eukaryotic cell contains two distinct genomes—the nuclear and the mitochondrial—whose coordinated interactions over billions of years now represent one of life's most important co-evolved mutualisms (Gillham 1994). Many gene products are encoded in the nucleus and exported to the mitochondria, where they interact with other, mitochondrially encoded, genes. Organismal fitness depends upon compatibility between nuclear and mitochondrial gene products (Meiklejohn et al. 2013), and these interactions (hereafter "mito-nuclear") are fundamental to eukaryotic existence and underlie key life history traits, including somatic maintenance, reproductive performance, and aging (Rand et al. 2004; Dowling et al. 2008).

However, because mitochondria are often maternally inherited, selection acting on these mito-nuclear interactions is asymmetrical in males and females. Mutations detrimental to males are not selected against unless they are also detrimental to females, except in some cases involving nonrandom mating, sperm limitation, or paternal mitochondrial transmission (e.g., Rand et al. 2001; Wade and Brandvain 2009; Unckless and Herren 2009; Hedrick 2011; Zhang et al 2012). In extreme cases, mitochondrial mutations that harm males can even be selected for if they benefit females. This results in a male mutational load, where mutations detrimental to males are not purged from populations and accumulate across generations (Frank and Hurst 1996; Gemmell et al 2004). This male mutational load can be detected in the form of male-biased gene mis-expression (Innocenti et al. 2011), reduction in male lifespan (Camus et al. 2012), and male fertility (Smith et al. 2010; Yee et al. 2013) in individuals that contain mitochondria from different populations.

Maternal inheritance of mitochondria puts mitochondrial genes in contrasting evolutionary dynamics with different parts of the nuclear genome: whereas Y chromosomes have strict paternal transmission, autosomes are equally transmitted through males and females, and X chromosomes spend twice their time in females compared with males. This sexual asymmetry across the genome might set the scene for intergenomic coadaptation or conflict. On the one hand, we expect beneficial gene combinations to be facilitated if genes that interact with the mitochondria are on the X chromosome. The X chromosomes in mammals and Drosophila have been shown to be feminized for gene expression (Khil et al. 2004; Meisel et al. 2012), and X-linked genes are co-transmitted with mitochondrial genes through the female two-third of the time. Under such a scenario—with inter-genomic co-adaptation driving the distribution of genes that interact with mitochondria—we might expect an over-representation of mito-nuclear genes on the X (Rand et al. 2001; Wade and Goodnight 2006; Brandvain and Wade 2009). On the other hand, the accumulation of mutations that are detrimental to males, referred to as male-biased mitochondrial mutational load, might be ameliorated if genes that interact with the mitochondria move off the X, onto parts of the genome with equal (or even male-biased) transmission. If conflict drives the distribution of mito-nuclear genes, we would expect an under-representation of genes that interact with the mitochondria on the X chromosome (Rice 1984; Werren 2011; Drown et al. 2012).

We might also expect converse patterns for Z chromosomes in female-heterogametic (ZW ZZ) species. ZW systems often show reverse patterns for sexual conflict scenarios because the Z is masculinized (Wright et al 2012) while the X is feminized for gene expression. This potentially results in an under-representation of mito-nuclear genes on the Z chromosome because mitochondria are co-transmitted with Z chromosomes only one-third of the time. Alternatively, because the Z and mitochondria can never be transmitted through males, it is possible that there is no expected bias on Z chromosomes with regard to mito-nuclear genes (Drown et al 2012). Finally, it has also been suggested that the Z chromosome might be enriched for mito-nuclear genes due to some types of sexual selection in males (Hill and Johnson 2013).

These predictions for the distribution of mitonuclear genes are predominantly based on probabilities of co-inheritance of mitochondria with different parts of the nuclear genome and do not take into account more complex processes such as linkage patterns of genes interacting with mitochondria. Empirical evidence for mito-sex chromosome interactions is not consistent. Some experimental evidence suggests genes on the X chromosome interact with mitochondrial genomes in *Drosophila* (Rand et al. 2001), whereas other assessments failed to detect mito-autosomal interactions (Clark 1985; Clark and Lyckegaard 1988). Consistent with the predictions of inter-genomic conflict, a strong under-representation of mitochondrial genes on the X chromosome was found across a range of mammal species (Drown et al. 2012). However, the data set used by Drown et al. (2012) is phylogenetically non-independent, as the X chromosomes in the therian mammals derived from the same common ancestor and show strong conservation of gene content across the clade (Veyrunes et al. 2008). Therefore, the broader generality of the dearth of mitochondrial genes on the X remains largely unexplored.

Here, we test the universality of predictions of mito-nuclear co-adaptation and conflict by exploring the genomic distribution of genes that interact with the mitochondrial genome. We extend previous studies by exploring these interactions on a broad scale, incorporating multiple examples of male- and female- heterogamety in species with independent origins of their sex chromosomes.

Materials and Methods

Detection and Localization of Genes Interacting with Mitochondria

In order to expand our analysis to species with less complete genome annotations, we modified the protocol from Drown et al. (2012) to compare the chromosomal distribution of genes that interact with the mitochondria across a range of species with phylogenetically independent sex chromosomes. In the first step, we obtained the proteomes for the several therian mammals (*Bos taurus, Pan troglodytes, Canis familiaris, Gorilla gorilla, Homo sapiens, Macaca mulatta, Equus caballus, Oryctolagus cuniculus, Pongo abelii, Rattus norvegicus, Sus scrofa*, and *Monodelphis domestica*), the monotreme *Ornithorhynchus anatinus*, three birds (*Gallus gallus*, *Meleagris gallopavo,* and *Taeniopygia guttata*), the stickleback fish *Gasterosteus aculeatus*, *Drosophila melanogaster,* and *Caenorhabditis elegans* from Ensembl v71 (Flicek et al. 2013). In order to increase the number of independently-evolved sex chromosomes, we also obtained the proteomes for *Tribolium castaneum*, *Bombyx mori,* and *Schistosoma mansoni* from Ensembl Metazoa v18 (Kersey et al. 2012).

Because genome and gene ontology (GO) annotation quality varies across our species, we used a reciprocal best BLAST hit (rBBH) approach to find one-to-one orthologs between the well-annotated *Mus musculus* mito–nuclear genes and the other species using the catalog of genes with mitochondrial annotation (mito-nuclear genes) in the GO (Ashburner et al. 2000) ID 0005739 for *M. musculus* using Biomart (Durinck et al. 2005) from Ensembl v71 (Flicek et al. 2013). This approach relies on the high level of conservation of mitochondrial gene function (Jafari et al. 2013; Lotz et al. 2014). To verify that rBBH is appropriate for mito–nuclear genes, we compared the list of genes obtained through rBBH with the list of mitochondrially annotated genes using GO term GO:0005739 in Biomart for *D. melanogaster* and *C. elegans*—two species with more complete gene annotation. We found that out of the 522 *D. melanogaster* GO:0005739 genes, 66% (345/522) were also identified as mito-nuclear by the rBBH. Of the 251 *C. elegans* GO:0005739 genes, only 7% (18/251) were identified through the rBBH. This suggests that, while rBBH is useful for detecting mito-nuclear orthologs (comparable with computational annotation of GO terms), our approach may miss or incorrectly classify some of the mito-nuclear genes across distantly related species.

In order to account for clade-specific differences, we conducted two further analyses. First, we repeated the rBBH analysis, using Biomart to identify mito-nuclear

GO:0005739 genes for *D. melanogaster* and *C. elegans* in addition to *M. musculus.* Because these are relatively well annotated genomes, we used them as clade-specific reference species in order to reduce taxonomic distance. Therefore, we used 1) *M. musculus* mito-nuclear genes as the reference for other vertebrates (Theria, *O. anatinus, G. aculeatus*, and Aves), 2) *D. melanogaster* mito-nuclear genes as the reference set for other insects (*T. castaneum* and *B. mori*), and 3) *C. elegans* mito-nuclear genes for the entozoans (with *S. mansoni*). Second, we also present results using just Biomart GO term annotations for those species where gene products have been annotated.

For the rBBH analysis, we used the longest protein isoform and only considered hits when the BLASTP (Altschul et al. 1997) e-value was below 10-$^{-7}$. In the second rBBH analysis, also using *D. melanogaster* and *C. elegans* as reference points, we used a more stringent e-value threshold of 10-$^{-10}$; hits were then ordered by bitscore, and an rBBH was accepted only when the best hit had a sequence identity larger than 30%. After the rBBH analyses, we determined the chromosomal location for mouse mito-nuclear orthologs in each species. The *S. mansoni* locations are based on Vicoso and Bachtrog (2011), *B. mori* positions were extracted from KAIKObase version 3.2.1 (Shimomura et al. 2009), *T. castaneum* are based on Ensembl Metazoa v18 (Kersey et al. 2012), and all other locations are based on Ensembl v71 (Durinck et al. 2005).

As a result, we created three lists of nuclear genes with mitochondrial annotation and their chromosomal locations: 1) using direct GO annotation (only in *M. musculus*) or based on orthology predictions (all other species), 2) based on direct GO annotation (*M. musculus*, *D. melanogaster* and *C. elegans*) or based on orthology predictions using the closest relative from these three species, and 3) based on direct GO annotation, just for *O. anatinus* and *G. aculeatus* (*S. mansoni*, *T. castaneum* and *B. mori* are not available in Ensembl, and Theria and Aves have previously been reported using this approach by Drown et al. 2012).

Statistical Analysis

In order to avoid problems with phylogenetic non-independence, we combined all species that share the same orthologous sex chromosome into a single data point (i.e., the therian mammals were grouped together, as were the birds). We then compared the density of mito-nuclear genes on the sex chromosomes and the autosomes relative to the expected gene density based on the total number of mitochondrial annotated genes. For *D. melanogaster,* each Muller element (X, 2L, 2R, 3L, 3R, 4) was treated as a separate chromosome. The expected gene count per chromosome was calculated as the total number of mito-nuclear genes multiplied by the proportion of all annotated genes on each chromosome. The bias of mito-nuclear genes was the ratio of the observed number of mito-nuclear genes on a chromosome to the expected count, where an over-representation is a bias $>$ 1 and an under-representation is a bias $<$ 1. In *G. aculeatus*, we also included the neo-sex chromosome (Kitano et al. 2009; Natri et al. 2013), as well as the *D. melanogaster* ancient-sex chromosome, which displays many properties of an X chromosome (Vicoso and Bachtrog 2013). The only sex-limited sex chromosome with sufficient size and annotation was the *S. mansoni* W, which is also included.

We tested the significance of the over- or under-representation of mitochondrial genes on the sex chromosomes by bootstrapping. To calculate confidence intervals (CIs) for sex chromosome bias, for each species/clade, we sampled with replacement 10,000 times the number of genes on the sex chromosome, summed the number of genes with mitochondrial annotation, calculated bias (as above) and took the 95% CIs of the distribution. To calculate CIs for bias on the autosomes, we sampled with replacement 1,000 times the genes on each of the autosomes (i.e., between 4 and 27 chromosomes, depending upon the clade), calculated bias for each chromosome, calculated the mean bias for each sampling event, and calculated the 95% CIs of the mean (i.e., the CI was calculated from 1,000 samples, and each sample was the mean bias of all chromosomes). For each analysis we corrected for multiple testing for nine different sex chromosomes, at an alpha of 0.05, using Bonferronni correction ($P < 0.0057$). Sex chromosomes had a significant over- or under-representation of mitochondrial genes if the sex chromosome CI did not overlap the CI of the autosomes.

When grouping different species together (the Theria, as well as Aves) or when one species has multiple sex chromosomes (*O. anatinus*), we calculated the CI for sex chromosome bias by summing together all the genes on the sex chromosomes and treating them as one large sex chromosome. When testing the autosomal distribution of the grouped species, sampling with replacement was done from each species such that each species contributed equally to the sampling distribution (i.e., to the 1,000 bootstrapped data points). We tested whether the bias of neo-, ancient-, and sex-limited chromosomes was different from the autosomes by bootstrapping all autosomal genes and excluding the homogametic sex chromosome.

We tested the significance of the overall over- or under-representation of mito-nuclear genes on the sex chromosomes in male- and female-heterogametic systems by bootstrapping 10,000 times the bias for each orthologous sex chromosome (mean bias for those sex chromosomes represented by multiple species) and calculating the 95% CIs for X and Z chromosomes. This slightly different approach to the previous bootstrapping technique enabled each clade to contribute equally to the distribution, irrespective of the size of the sex chromosome.

The significance of over- or under-representations of mito-nuclear genes on the sex chromosomes were also analyzed using χ^2 tests.

Results and Discussion

It has been previously suggested that the paucity of mito-nuclear genes on the therian X chromosome was driven by sexual conflict related to asymmetrical inheritance (Drown et al. 2012). Mito-nuclear genes have been suggested to move off the X onto autosomes due to conflict between the sexes, a process that involves gene duplication, fixation, followed by loss of the sex-chromosome linked parent copy (Gallach et al 2012; Drown et al. 2012). Genes with effects that can ameliorate male-detrimental mitochondrial mutations would be selected in males and are more likely to accumulate on autosomes than on female-biased X chromosomes. Although some have suggested that there should be a random distribution of mito-nuclear genes on Z chromosomes (Drown et al. 2012), others have predicted an over-representation of mito-nuclear genes on the Z chromosome of female heterogametic species related to sexual selection (Hill and Johnson 2013).

If sexual conflict over asymmetrical inheritance does shape the distribution of mito-nuclear genes, we might expect convergent patterns of under-representation across independent X chromosomes (Drown et al. 2012). X chromosomes have in general fewer mito-nuclear genes (i.e., bias < 1) than expected (mean bias = 0.86, CI = 0.72—1.00); however, only two of six independent X chromosomes showed statistically significant under-representations of mito-nuclear genes. The therian mammals exhibit the most extreme distribution of mito-nuclear genes on the X chromosome, with only the *C. elegans* X chromosome showing a significant paucity.

Furthermore, *C. elegans* is a gynodioecios species, with both males and hermaphrodites. The lack of distinct male and female individuals within the species may limit the degree of sexual conflict, as male-harming mutations in mito-nuclear genes would also affect the male function in hermaphrodites. This suggests that sexual conflict may be reduced in this species and may not be the driver of the distribution of mito-nuclear genes. However, it is important to note that gynodioecy is a recently derived trait in the *Caenorhabditis* lineage, and most other species in the genus are fully gonochoristic. This means that any reduction in sexual conflict due to gynodioecy would have been relatively recent.

We also explored the neo-X chromosome in *G. aculeatus* (Kitano et al. 2009; Natri et al. 2013) and the B chromosome in *D. melanogaster*, which has recently been shown to be an ancient sex chromosome that has reverted to an autosome in the *Drosophila* lineage (Vicoso and Bachtrog 2013), in order to test whether recent and past evolutionary history shape current patterns. Both the *G. aculeatus* X and neo-X showed no significant bias of mito-nuclear genes (tables 1–3). The ancient X chromosome in *D. melanogaster* also showed no overall bias (tables 1 and 2).

These results across multiple independent X chromosomes suggest that patterns of mito-nuclear gene distribution are not consistently shaped by convergent sexual conflict over asymmetrical inheritance across independent sex chromosome systems. This pattern was consistent across all rBBH approaches (figs. 1 and 2, tables 1 and 2) and species-specific GO annotations (fig. 3 and table 3).

Many patterns driven by sexual conflict on X chromosomes are predicted to display converse patterns on Z chromosomes (Rice 1984), and this has been true for genomic characters including the sexualization of gene expression (Dean and Mank 2014). We might therefore also expect convergent over-representation of mito-nuclear genes on Z chromosomes, although the low co-transmission between the mitochondria and the Z chromosome may ameliorate this prediction (Drown et al. 2012). Our results indicate that Z chromosomes overall have slightly more mito-nuclear genes (i.e., bias > 1) than expected (mean bias = 1.06, CI = 1.02—1.11), but there was no taxon-specific case where a Z chromosome carried a significantly greater proportion of mito-nuclear genes than expected based on its relative size.

The W chromosome and mitochondria are in complete linkage, being co-transmitted each generation. Consequently, we may expect an over-representation of co-adapted, female-benefitting mito-nuclear genes on the W. Although we do observe some W-linked mito-nuclear genes in *S. mansoni*, suggesting that some genes have sex-specific expression, there is no significant over-representation of these genes on the W chromosome (tables 1 and 2). The lack of bias of mito-nuclear genes on W could be due to lack of selection for gene movement in the female—the mitochondria is already optimized for females and so no advantage for the female is gained by movement of Z or autosomal genes onto the W.

It is possible that the genomic distribution of mito-nuclear genes is somewhat confounded by other genomic phenomena. First, mitochondrial mutation rate differs substantially across species; for example, mammals tend to have high rates and *Drosophila* have low rates (Montooth and Rand 2008). Mitochondrial mutation rate will affect the extent to which mitochondria can evolve female-beneficial mutations. Second, the relative rate of evolution of sex chromosomes to autosomes (the Faster-X Effect, Charlesworth et al. 1987) varies across species and depends on the relative effective population size of the X compared with the autosomes (Mank et al. 2010). The relative effective population size of different X chromosomes to autosomes varies substantially (Mank et al. 2010 and references therein); however, this does not necessarily explain our data, as, for example, *E. caballus* and *D. melanogaster* both have high relative effective population sizes of the X chromosome (Andolfatto 2001; Connallon 2007; Singh et al. 2007; Lau et al. 2009), and yet *D. melanogaster* shows no overall bias, while *E. caballus*

Table 1
Mean Bias and 95% CIs of Mito-nuclear Genes on the Sex Chromosomes and Autosomes

Species or Clade	Over-/Under-representation of Mito-nuclear Genes on Sex Chromosome (Bias)	95% Bonferroni-Corrected CI of the Sex Chromosome	95% Bonferonni-Corrected CI of the Autosomes	χ^2 Test and *P* Value
Male heterogamety	0.86	0.72–1.00		
Therian mammals	**Under (mean = 0.64)**	**0.55–0.72**	**0.90–1.13**	**89.5, *P* < 0.0001**
H. sapiens	0.63			
P. troglodytes	0.69			
G. gorilla	0.62			
P. abelii	0.60			
M. mulatta	0.65			
E. caballus	0.59			
B. taurus	0.64			
S. scrofa	0.77			
O. cuniculus	0.63			
R. norvegicus	0.60			
M. musculus	0.69			
M. domestica	0.44			
O. anatinus	Under (mean = 0.85)	0.45–1.26	0.64–1.27	0.92, *P* = 0.34
G. aculeatus	Under (0.88)	0.57–1.20	0.92–1.09	0.93, *P* = 0.33
D. melanogaster	Over (1.11)	0.89–1.33	0.77–1.23	2.17, *P* = 0.14
T. castaneum	Over (1.06)	0.69–1.42	0.91–1.11	0.18, *P* = 0.67
C. elegans	**Under (0.72)**	**0.51–0.92**	**0.98–1.18**	**12.06, *P* = 0.0005**
Female heterogamety	1.06	1.02–1.11		
Aves	Over (mean = 1.07)	0.86–1.28	0.86–1.09	0.92, *P* = 0.34
G. gallus	1.10			
M. gallopavo	0.97			
T. guttata	1.12			
B. mori	Over (1.02)	0.61–1.43	0.86–1.04	0.01, *P* = 0.90
S. mansoni	Over (1.11)	0.61–1.60	0.87–1.17	0.41, *P* = 0.52
Sex-limited/neo/ancient				
G. aculeatus neo-X	Under (0.92)	0.57–1.19	0.92–1.09	0.47, *P* = 0.59
D. melanogaster ancient-X (chromosome 4)	1.00	−0.08–2.08	0.91–1.09	0.00, *P* = 0.97
S. mansoni W	Under (0.90)	0.63–1.16	0.85–1.18	1.00, *P* = 0.32

NOTE.—Significant under or over-representations are in bold. CIs calculated by bootstrapping. χ^2 statistics are also presented. One-to-one orthologs were identified using *M. musculus* as the reference.

shows an under-representation (tables 1 and 2). Third, we may expect variation in the magnitude of the male-biased mutation rate, for example, due to species differences in generation time and in the strength of sexual selection and associated intensity of sperm competition (Ellegren 2007). However, it is difficult to see how the patterns we observe are driven by variation in male-biased mutation. Finally, levels of gene transfer and genome rearrangement are lineage-specific (Rand et al. 2001), where low levels of movement will restrict the ability of different parts of the genome to respond to inter-genomic coadaptation and conflict. This may explain many of the non-significant associations.

Alternatively, interactions between the mitochondrial genome and the X and Z chromosome have been suggested to play a role in sexual selection and might be enriched for mito-nuclear genes that play a role in coloration, such as those involving carotenoids (Hill and Johnson 2013). We did not observe this predicted over-representation on any Z chromosomes, and it is difficult to see how differences among our study species in the degree and type of sexual selection explain the variance in the distribution of mitochondrial genes.

A further possibility is that the genomic distribution of mito-nuclear genes is driven by gametic function. Although mitochondrial activity is generally not crucial for non-motile egg function (de Paula et al. 2013), it is integral to sperm energy production and motility (Cummins 2009). Although many genes are functionally diploid in sperm (Braun et al. 1989), there is evidence that many genes are expressed within the spermatid and are subject to haploid selection (Joseph and Kirkpatrick 2004). Because any single spermatozoon will only carry either an X or Y chromosome, expression of mito-nuclear genes within the sperm would lead to selection against sex- linkage as half of the male gametes would lack

Table 2

Mean Bias and 95% CIs of Mito-nuclear Genes on the Sex Chromosomes and Autosomes

Species or Clade	Over-/Under-representation of Mito-nuclear Genes on Sex Chromosome (Bias)	95% Bonferroni-Corrected CI of the Sex Chromosome	95% Bonferonni-Corrected CI of the Autosomes	χ^2 Test and *P* Value
Male heterogamety				
Therian mammals	**Under (mean = 0.71)**	**0.61–0.79**	**0.90–1.13**	**62.8, *P* < 0.0001**
H. sapiens	0.73			
P. troglodytes	0.69			
G. gorilla	0.72			
P. abelii	0.69			
M. mulatta	0.72			
E. caballus	0.64			
B. taurus	0.71			
S. scrofa	0.87			
O. cuniculus	0.77			
R. norvegicus	0.65			
M. musculus	0.68			
M. domestica	0.48			
O. anatinus	Under (mean = 0.83)	0.43–1.22	0.69–1.29	1.38, *P* = 0.24
G. aculeatus	Under (0.92)	0.60–1.23	0.93–1.09	0.47, *P* = 0.49
D. melanogaster	No bias (1.00)	0.70–1.30	0.86–1.13	0.00, *P* = 0.99
T. castaneum	Under (0.96)	0.37–1.55	0.84–1.14	0.03, *P* = 0.86
C. elegans	**Under (0.23)**	**0.0–0.46**	**0.91–1.28**	**23.8, *P* < 0.0001**
Female heterogamety				
Aves	Over (mean = 1.02)	0.83–1.22	0.86–1.09	0.10, *P* = 0.75
G. gallus	1.06			
M. gallopavo	0.89			
T. guttata	1.10			
B. Mori	Under (0.84)	0.22–1.45	0.83–1.12	0.47, *P* = 0.49
S. Mansoni	Under (0.52)	−0.50–1.54	0.64–1.69	0.95, *p* = 0.33
Sex-limited/neo/ancient				
G. aculeatus neo-X	Under (0.84)	0.54–1.13	0.92–1.09	1.96, *P* = 0.16
D. melanogaster ancient-X (chromosome 4)	Under (0.99)	−0.58–2.55	0.86–1.13	0.00, *P* = 0.99
S. mansoni W	Under (1.04)	0.18–1.90	0.61–1.77	0.00, *P* = 0.97

Note.—Significant under or over-representations are in bold. CIs calculated by bootstrapping. Mito-nuclear genes detected by the rBBH analysis using *M. musculus, D. melanogaster,* and *C. elegans* to find orthologs.

Table 3

Mean Bias and 95% CIs of Mito-nuclear Genes on the Sex Chromosomes and Autosomes

Species or Clade	Over/underrepresentation of Mitonuclear Genes on Sex Chromosome (Bias)	95% Bonferroni-Corrected CI of the Sex Chromosome	95% Bonferroni-Corrected CI of the Autosomes	χ^2 Test and *P* Value
Male heterogamety				
O. Anatinus	Under (mean = 0.87)	0.41–1.33	0.36–1.35	0.60, *P* = 0.44
G. Aculeatus	Under (0.34)	−0.58–1.23	0.66–1.44	1.46, *P* = 0.23
Sex-limited/neo/ancient				
G. aculeatus neo-X	1.00	−0.61–2.60	0.63–1.44	0.00, *P* = 0.95

Note.—Mitonuclear genes identified using GO terms in Biomart.

a functional copy. Conversely, all sperm in female heterogametic species contain a Z chromosome, and there would be no expected selection against Z-linkage of mito-nuclear genes.

Furthermore, differences among taxa in sperm biology could explain some of the patterns we observe among male heterogametic taxa. For example, species differ in the presence or absence of sperm hyper-activation, which requires

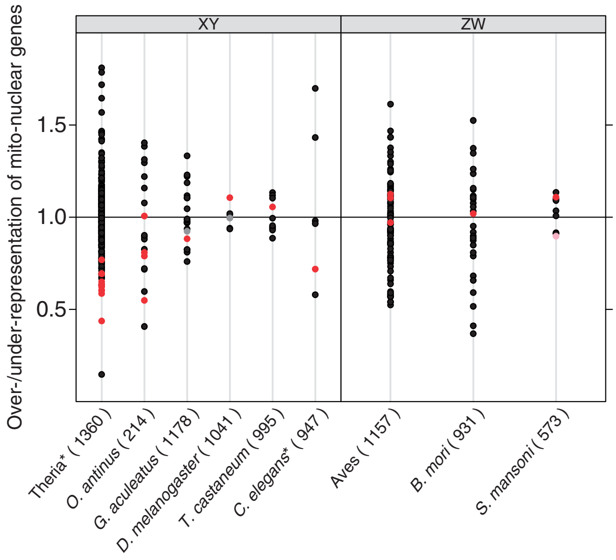


Fig. 1.—Bias of nuclear–mitochondrial genes on the sex chromosomes across species with independent sex chromosomes. Values for each autosome are in black, major sex chromosomes (X or Z) in red, old (i.e., *D. melanogaster* fourth) and neo (i.e., *G. aculeatus* chromosome 9) in gray, and the *S. mansoni* W chromosome in pink. Values in parenthesis after species names indicate the total number of mito-nuclear genes in the genome detected by the rBBH analysis with *M. musculus*. Species marked by * have a significant under-representation of nuclear–mitochondrial genes on the X chromosome. Note: Some of *D. melanogaster* autosomal points overlap.

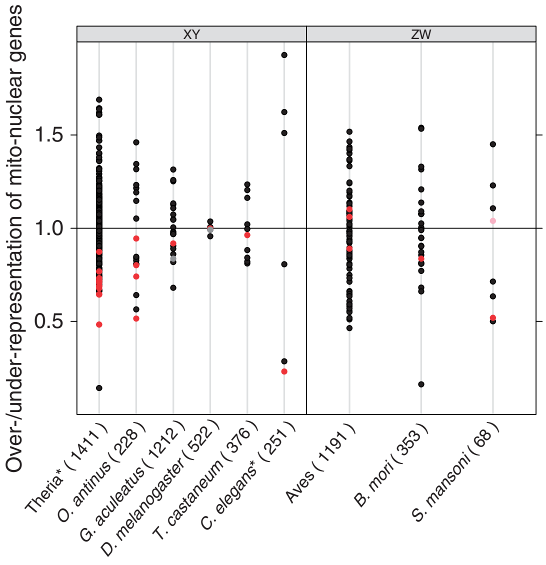


Fig. 2.—Bias of nuclear–mitochondrial genes on the sex chromosomes across species with independent sex chromosomes. Values for each autosome are in black, major sex chromosomes (X or Z) in red, old (i.e., *D. melanogaster* fourth) and neo (i.e., *G. aculeatus* chromosome 9) in gray, and the *S. mansoni* W chromosome in pink. Values in parenthesis after species names indicate the total number of mitonuclear genes in the genome detected by the rBBH analysis using *M. musculus, D. melanogaster,* and *C. elegans* to find orthologs. Species marked by * have a significant underrepresentation of nuclear–mitochondrial genes on the X chromosome. Note: Some of *D. melanogaster* autosomal points overlap.

high mitochondrial activity (Cummins 2009). Also, the degree to which oxidative metabolism is required for sperm motility differs, and both human and mouse sperm do not need mitochondrial activity for motility (Cummins 2009). Factors such as this may affect the degree of haploid expression of mito-nuclear genes in sperm and therefore the distribution of mito-nuclear genes on X chromosomes. However, we hasten to point out that none of these explanations alone fully account for why Theria and *C. elegans* have an under-representation of mito-nuclear genes on their X chromosomes. More complex theory, taking into account patterns of gene duplication and gene movement, may be required to make sense of these patterns.

The need to maximize the number of independent sex chromosomes in our analyses means that we had to include some genomes with incomplete functional annotation. To solve this, we employed an rBBH approach in order to detect orthologs of mitochondrial interacting genes that are annotated in model organisms like *M. musculus*, *D. melanogaster,* and *C. elegans*. However, this approach could be influenced by taxon-specific mito-nuclear genes and difficulties in orthology identification across large evolutionary distances. Although this does limit the number of genes we identify through strict orthology identification in some taxa, we do not believe that it has unduly biased our results for several reasons. First, nuclear genes that interact with the mitochondria are conserved across broad taxonomic groups (Porcelli et al. 2007; Lotz et al. 2014), suggesting that rBBH is broadly applicable. The convergence between our results using *M. musculus* as the reference for all rBBH with results using *D. melanogaster* and *C. elegans* as reference suggests that conservation predominates over clade- or species-specific patterns. We also detected similar patterns using species-specific GO annotations.

In conclusion, our results are not universally consistent with either sexual conflict (Drown et al. 2012) or sexual selection (Hill 2013; Hill and Johnson 2013), driving the general distribution of mito-nuclear genes on all sex chromosomes. We observed significant under-representation of mito-nuclear genes in just two of six analyzed X chromosomes, and no patterns of non-random distribution on any analyzed Z chromosome. The results suggest that other genomic phenomena may limit the extent to which inter-genomic conflict (Drown et al. 2012) or sexual selection (Hill and Johnson 2013) affect mito-nuclear distributions and confirm the importance of broad, phylogenetically independent analysis.

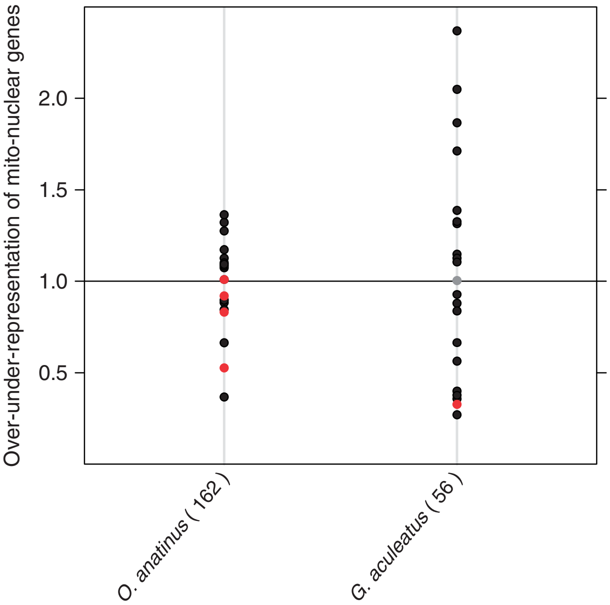


Fig. 3.—Bias of nuclear–mitochondrial genes on the sex chromosomes across *G. aculeatus* and *O. anatinus*. Values for each autosome are in black, X chromosomes in red, and neo (i.e. *G. aculeatus* chromosome 9) in gray. Values in parenthesis after species names indicate the total number of mitonuclear genes in the genome detected using GO:0005739 to identify genes that interact with the mitochondria.

Acknowledgments

We thank Beatriz Vicoso and Doris Bachtrog for data on Schistosoma mansoni X chromosome contigs and Michael Stocks for computational advice. This work was supported by grants from the Biotechnology and Biological Sciences Research Council and European Research Council (grant agreement 260233), a short-term fellowship at the Wissenschaftskolleg zu Berlin to J.E.M., and a Helge Ax:son Johnsons stipend to R.D.

Literature Cited

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389–3402.

Andolfatto P. 2001. Contrasting patterns of X-linked and autosomal nucleotide variation in *Drosophila*. Mol Biol Evol. 18: 279–290.

Ashburner M, et al. 2000. Gene Ontology: tool for the unification of biology. Nat Genet. 25:25–29.

Brandvain Y, Wade MJ. 2009. The functional transfer of genes form the mitochondria to the nucleus: the effects of selection, mutation, population size and rate of self fertilization. Genetics 182: 1129–1139.

Braun RE, Behringer RR, Peschon JJ, Brinster RL, Palmiter RD. 1989. Genetically haploid spermatids are phenotypically diploid. Nature 337:373–376.

Camus MF, Clancy DJ, Dowling DK. 2012. Mitochondria, maternal inheritance, and male aging. Curr Biol. 22:1717–1721.

Charlesworth B, Coyne JA, Barton NH. 1987. The relative rates of evolution of sex chromosomes and autosomes. Am Nat. 130: 113–146.

Clark AM. 1985. Natural selection with nuclear and cytoplasmic transmission. II. Tests with *Drosophila* from diverse populations. Genetics 111: 97–112.

Clark AM, Lyckegaard EMS. 1988. Natural selection with nuclear and cytoplasmic transmission. III. Joint analysis of segregation and mitochondrial DNA in *Drosophila melanogaster*. Genetics 118: 471–481.

Connallon T. 2007. Adaptive protein evolution of X-linked and autosomal genes in *Drosophila*. Mol Biol Evol. 24:2566–2572.

Cummins J. 2009. Sperm motility and energetics. In: Birkhead TR, Hosken DJ, Pitnick S, editors. Sperm biology. Academic Press, London. p. 185–206.

Dean R, Mank JE. Forthcoming. 2014. The role of sex chromosomes in sexual dimorphism: discordance between molecular and phenotypic data. J Evol Biol. Advance Access published February 21, 2014, doi:10.1111/jeb.12345.

de Paula WBM, et al. 2013. Female and male gamete mitochondria are distinct and complementary in transcription, structure and genome function. Genome Biol Evol. 5:1969–1977.

Dowling DK, Friberg U, Lindell J. 2008. Evolutionary implications of non-neutral mitochondrial genetic variation. Trends Ecol Evol. 23: 546–554.

Drown DM, Preuss KM, Wade MJ. 2012. Evidence of a paucity of genes which interact with the mitochondrion on the X in mammals. Genome Biol Evol. 4:763–768.

Durinck S, et al. 2005. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics 21: 3439–3440.

Ellegren H. 2007. Characteristics, causes and evolutionary consequences of male-biased mutation. Proc Roy Soc B. 274:1–10.

Flicek P, et al. 2013. Ensembl 2013. Nucleic Acids Res. 41:D48–D55.

Frank SA, Hurst LD. 1996. Mitochondria and male disease. Nature 383: 224.

Gallach M, Chandrasekaran C, Betran E. 2012. Analyses of nuclearly encoded mitochondrial genes suggest gene duplication as a mechanism for resolving intralocus sexually antagonistic conflict in *Drosophila*. Genome Biol Evol. 2:835–850.

Gemmell NJ, Metcalf VJ, Allendorf FW. 2004. Mother's curse: the effect of mtDNA on individual fitness and population viability. Trends Ecol Evol. 19:238–244.

Gillham NW. 1994. Organelle genes and genomes. New York: Oxford University Press.

Hedrick PW. 2011. Reversing mothers curse revisited. Evolution 66: 612–616.

Hill GE. 2014. Sex linkage of nuclear-encoded mitochondrial genes. Heredity. 112:469–470.

Hill GE, Johnson JD. 2013. The mitonuclear compatibility hypothesis of sexual selection. Proc Roy Soc B. 280:20131314.

Innocenti P, Morrow EH, Dowling DK. 2011. Experimental evidence supports a sex-specific selective sieve in mitochondrial genome evolution. Science 332:845–848.

Jafari M, Sadeghi M, Mirzaie M, Marashi S, Rezaei-Tavirani M. 2013. Evolutionarily conserved motifs and modules in mitochondrial protein-protein interaction networks. Mitochondrion 13:668–75.

Joseph SB, Kirkpatrick M. 2004. Haploid selection in animals. Trends Ecol Evol. 19:592–597.

Kersey PJ, et al. 2012. Ensembl genomes: an integrative resource for genome-scale data from non-vertebrate species. Nucleic Acids Res. 40:D91–D97.

Khil PP, Smirnova NA, Romanienko PJ, Camerini-Otero RD. 2004. The mouse X chromosome in enriched for sex-biased genes not subject to selection by meiotic sex chromosome inactivation. Nat Genet. 36:642–646.

Kitano J, et al. 2009. A role for a neo-sex chromosome in stickleback speciation. Nature 461:1079–1083.

Lau A, Peng L, Goto H, Chemnick L, Ryder O, Makova K. 2009. Horse domestication and conservation genetics of Przewalski's horse inferred from. Mol Biol Evol. 26:199–208.

Lotz C, et al. 2014. The characterization, design, and function of the mitochondrial proteome: from organs to organisms. J Proteome Res. 13: 433–446.

Mank J, Vicoso B, Berlin S, Charlesworth B. 2010. Effective population size and the Faster-X effect: empirical results and their interpretation. Evolution 64:663–674.

Meiklejohn CD, et al. 2013. An incompatability between a mitochondrial tRNA and its nuclear-encoded tRNA synthetase compromises development and fitness in *Drosophila*. PLoS Genet. 9:e1003238.

Meisel RP, Malone JH, Clark AG. 2012. Disentangling the relationship between sex-biased gene expression and sex-linkage. Genome Res. 22:1255–1256.

Montooth K, Rand D. 2008. The spectrum of mitochondrial mutation differs across species. PLoS Biol. 6:e213.

Natri H, Shikano T, Merila J. 2013. Progressive recombination suppression and differentiation in recently evolved neo-sex chromosomes. Mol Biol Evol. 30:1131–1144.

Porcelli D, Barsanti P, Pesole G, Caggese C. 2007. The nuclear OXPHOS genes in insecta: a common evolutionary origin, a common cis-regulatory motif, a common destiny for gene duplicates. BMC Evol Biol. 7: 215.

Rand DM, Clark AM, Kann LM. 2001. Sexually antagonistic cytonuclear fitness interactions in *Drosophila melanogaster*. Genetics 159: 173–187.

Rand DM, Haney RA, Fry AJ. 2004. Cytonuclear coevolution: the genomics of cooperation. Trends Ecol Evol. 19:645–653.

Rice WE. 1984. Sex chromosomes and the evolution of sexual dimorphism. Evolution 38:735–742.

Shimomura M, et al. 2009. KAIKObase: an integrated silkworm genome database and data mining tool. BMC Genomics 10:486.

Singh N, Macpherson J, Jensen J, Petrov D. 2007. Similar levels of X-linked and autosomal nucleotide variation in African and non-African populations of *Drosophila melanogaster*. BMC Evol Biol. 7:202.

Smith S, Turbill C, Suchentrunk F. 2010. Introducing mother's curse: low male fertility associated with an imported mtDNA haplotype in a captive colony of brown hares. Mol Ecol. 19:36–43.

Unckless RL, Herren JK. 2009. Population genetics of sexually antagonistic mitochondrial mutants under inbreeding. J Theor Biol. 260:132–136.

Veyrunes F, et al. 2008. Bird-like sex chromosomes of platypus imply recent origin of mammal sex. Genome Res. 18:965–973.

Vicoso B, Bachtrog D. 2011. Lack of global dosage compensation in *Schistosoma mansoni*, a female-heterogametic parasite. Genome Biol Evol. 3:230–235.

Vicoso B, Bachtrog D. 2013. Reversal of an ancient sex chromosome to an autosome in *Drosophila*. Nature 499:332–335.

Wade MJ, Brandvain Y. 2009. Reversing mother's curse: selection on male mitochondrial fitness. Evolution 63:1084–1089.

Wade MJ, Goodnight CJ. 2006. Cyto-nuclear epistasis: two-locus random genetic drift in hermaphroditic and dioecious species. Evolution 60: 643–659.

Werren JH. 2011. Selfish genetics elements, genetic conflict and evolutionary innovation. Proc Natl Acad Sci U S A. 108: 10863–10870.

Wright AE, Moghadam HK, Mank JE. 2012. Trade-off between selection for dosage compensation and masculinization of the avian Z chromosome. Genetics 192:1433–1445.

Yee WKW, Sutton KL, Dowling DK. 2013. In vivo male fertility is affected by naturally occurring mitochondrial haplotypes. Curr Biol. 23: R55–R56.

Zhang H, Guillaume F, Engelstadter J. 2012. The dynamics of mitochondrial mutations causing male infertility in spatially structured populations. Evolution 66:3179–3188.

Associate editor: Marta Wayne

The following article was first published in

Genome Biology and Evolution

Dean R, **Zimmer F**, Mank JE (2015) Deficit of mito-nuclear genes on the human X chromosome predates sex chromosome formation. *Genome Biology and Evolution* 7(2):636-641, **doi**: 10.1093/gbe/evv017

Deficit of Mitonuclear Genes on the Human X Chromosome Predates Sex Chromosome Formation

Rebecca Dean*, Fabian Zimmer, and Judith E. Mank

Department of Genetics, Evolution and Environment, University College London, United Kingdom

*Corresponding author: E-mail: r.dean@ucl.ac.uk.

Accepted: January 22, 2015

Abstract

Two taxa studied to date, the therian mammals and *Caenorhabditis elegans*, display underrepresentations of mitonuclear genes (mt-N genes, nuclear genes whose products are imported to and act within the mitochondria) on their X chromosomes. This pattern has been interpreted as the result of sexual conflict driving mt-N genes off of the X chromosome. However, studies in several other species have failed to detect a convergent biased distribution of sex-linked mt-N genes, leading to questions over the generality of the role of sexual conflict in shaping the distribution of mt-N genes. Here we tested whether mt-N genes moved off of the therian X chromosome following sex chromosome formation, consistent with the role of sexual conflict, or whether the paucity of mt-N genes on the therian X is a chance result of an underrepresentation on the ancestral regions that formed the X chromosome. We used a synteny-based approach to identify the ancestral regions in the platypus and chicken genomes that later formed the therian X chromosome. We then quantified the movement of mt-N genes on and off of the X chromosome and the distribution of mt-N genes on the human X and ancestral X regions. We failed to find an excess of mt-N gene movement off of the X. The bias of mt-N genes on ancestral therian X chromosomes was also not significantly different from the biases on the human X. Together our results suggest that, rather than conflict driving mt-N genes off of the mammalian X, random biases on chromosomes that formed the X chromosome could explain the paucity of mt-N genes in the therian lineage.

Key words: sexual conflict, sex chromosomes, mitochondria, synteny.

Introduction

A series of studies have recently generated substantial debate over the role of intergenomic conflict in driving mitonuclear (mt-N) gene distributions on and off sex chromosomes (Drown et al. 2012; Hill and Johnson 2013; Dean et al. 2014; Hough et al. 2014; Rogell et al. 2014). mt-N genes are loci whose products, encoded by the nuclear genome, are then imported into the mitochondria, which is the primary site of their activity. Because mitochondria and sex chromosomes have different inheritance patterns between the sexes, intergenomic conflict has been suggested as a potential explanation for the underrepresentation of mt-N genes on the X chromosomes of some animals (Drown et al. 2012; Dean et al. 2014). Mitochondria are maternally inherited in many species (although low rates of male transmission may occur, e.g., Wolff et al. 2013), and are therefore selected for female fitness effects, as male mitochondria are generally evolutionary dead ends. It has been shown that maternal transmission of mitochondria can result in quite serious costs to males, through the disruption of male function (Partridge and Hurst 1998; Innocenti et al. 2011; Drown et al. 2012).

The accumulation of mutations that are detrimental to males could be ameliorated if genes that interact with the mitochondria move to a more favorable genomic location for the evolution of compensatory mechanisms. Genes on the X chromosome, which spend two-thirds of their time in females, are more often cotransmitted with mitochondria than autosomal genes (Rand et al. 2001), and the X chromosome is also feminized in several species (reviewed in Dean and Mank 2014). This might make the X chromosome particularly unfavorable for male-biased compensation of the mitochondrial mutational load. It is therefore possible that there has been selection in males for the movement of mt-N genes off of the X chromosome in order to reduce disruption to male function induced by maternally transmitted mitochondria.

Consistent with the conflict hypothesis, *Caenorhabditis elegans* (Dean et al. 2014) and the therian mammals (Drown et al. 2012) show a deficit of mt-N genes on their X chromosomes, and genes sensitive to mitochondrial polymorphism are scarce on the *Drosophila* X chromosome (Rogell et al. 2014). However, a broader phylogenetic assessment of mt-N gene distributions revealed a mixed pattern, with most male heterogametic species studied showing no significant bias (Dean et al. 2014; Hough et al. 2014). Moreover, many sex-specific evolutionary properties observed on the X chromosome are observed in converse on Z chromosomes, such as distributions of sex-biased genes (Arunkumar et al. 2009; Wright et al. 2012), so we might expect a corresponding overabundance of Z-linked mt-N genes in female heterogametic systems; however, no such overabundance has yet been observed (Dean et al. 2014). Furthermore, if conflict is at least partly responsible for the genomic distribution of mt-N genes, it might also be expected to shape the distribution of nuclear genes that interact with the chloroplast, which is also often maternally inherited, but no bias was detected in the distribution of chloro-nuclear genes on the X chromosome in *Rumex* (Hough et al. 2014), a dioecious plant with sex chromosomes.

These patterns of mt-N gene distributions suggest that either conflict is particularly strong only in therian mammals and nematodes, or that some effect other than conflict explains the distribution in these two clades. The incorporation of mitochondrial loci into the nuclear genome began long before the formation of sex chromosomes in any single extant lineage (Dyall et al. 2004; Timmis et al. 2004; Cortez et al. 2014) and strong chromosomal biases exist for many autosomes, presumably due to chance variation in gene content (Drown et al. 2012; Dean et al. 2014; Hough et al. 2014). This presents the possibility that biases in mt-N gene distributions need not be driven by conflict, but instead could predate the formation of the sex chromosome, if the precursor autosomes showed an ancestral bias through chance alone.

We tested whether ancestral gene distributions can explain the underrepresentation of mt-N genes on therian sex chromosomes. The rapid gene and genome evolution in *Caenorhabditis* (Lipinski et al. 2011) precludes reconstruction of syntenic relationships across even closely related species, but amniotes have strongly conserved synteny (Dehal and Boore 2005), making it possible to identify syntenic regions in divergent taxa. In order to determine whether the paucity of mt-N genes on the therian X chromosome is a consequence of intergenomic sexual conflict, or whether it is simply the product of a biased distribution on the ancestral autosome that gave rise to the therian X chromosome, we tested the mt-N gene distributions on the ancestral regions syntenic to the therian X in platypus and chicken (hereafter termed X-syntenic regions).

We used the human X chromosome as our point of reference because of its excellent annotation. As the human X is broadly syntenic across therian mammals (Ohno 1967; Murphy et al. 1999; Band et al. 2000; Raudsepp et al. 2004), it is representative of the therian X in general. We identified regions in synteny with the human X in platypus (*Ornithorhynchus anatinus*) and chicken (*Gallus gallus*), the most recent ancestors to the Theria with different sex chromosomal systems (Graves 2006) and annotated genomes. This enabled us to use two complementary approaches to test the role of conflict in driving mt-N gene distributions. First, we identified orthologous genes, in platypus and chicken, to the human mt-N genes. We then tested for an excess of mt-N gene movement in order to investigate whether intergenomic conflict has driven mt-N genes off of the human X following sex chromosome formation. Second, we used these orthologous genes to compare mt-N gene distributions on human X and X-syntenic regions in platypus and chicken. If the abundance of mt-N genes on the X-syntenic regions is more than the abundance on the human X, then intergenomic conflict may have driven mt-N genes off of the therian X following sex chromosome formation. If, on the other hand, mt-N biases on the ancestral autosomes that gave rise to the therian X chromosome show a similar underrepresentation to the human X, then the chromosomal bias is unlikely to be a consequence of intergenomic conflict and may simply be a result of random variation across chromosomes in mt-N content.

Results and Discussion

mt-N Gene Movement On and Off the Human X Chromosome

We identified platypus chromosome 6 plus ten unmapped ultracontigs (platypus hX-syntenic regions), and regions of chicken chromosomes 1, 3, 4 and 12 (chicken hX-syntenic regions), as syntenic with the human X chromosome (fig. 1). The platypus hX-syntenic regions comprised a total of 381 genes spanning 71% of the length of the human X-chromosome and the chicken hX-syntenic regions comprised a total of 908 genes spanning 89% of the length of the human X-chromosome (fig. 1). The reduced coverage of the human X chromosome in platypus is largely due to the poorer assembly of the platypus genome.

To test whether an excess of mt-N gene movement off of the human X chromosome occurred following human X chromosome formation, we identified the location of the human mt-N orthologs in platypus and chicken. Pairs of orthologous genes that did not fall within syntenic blocks were potential candidates for genes that have moved. We identified four genes that moved onto the human X from Ultra contigs that were not in platypus hX-syntenic regions (from UltraContig 369; UltraContig 98; and two genes from UltraContig 519) and no genes that might have moved off the human X. These numbers were not significantly different than what we would expect based on the relative size and content of the X chromosome (Betrán et al. 2002; Vibranovski et al. 2009; Toups et al. 2011; Fisher’s exact test, $P > 0.6$), suggesting no excess of gene movement onto or off of the human X chromosome (table 1*a*). However, two of the genes that might have moved onto the X were from UltraContig

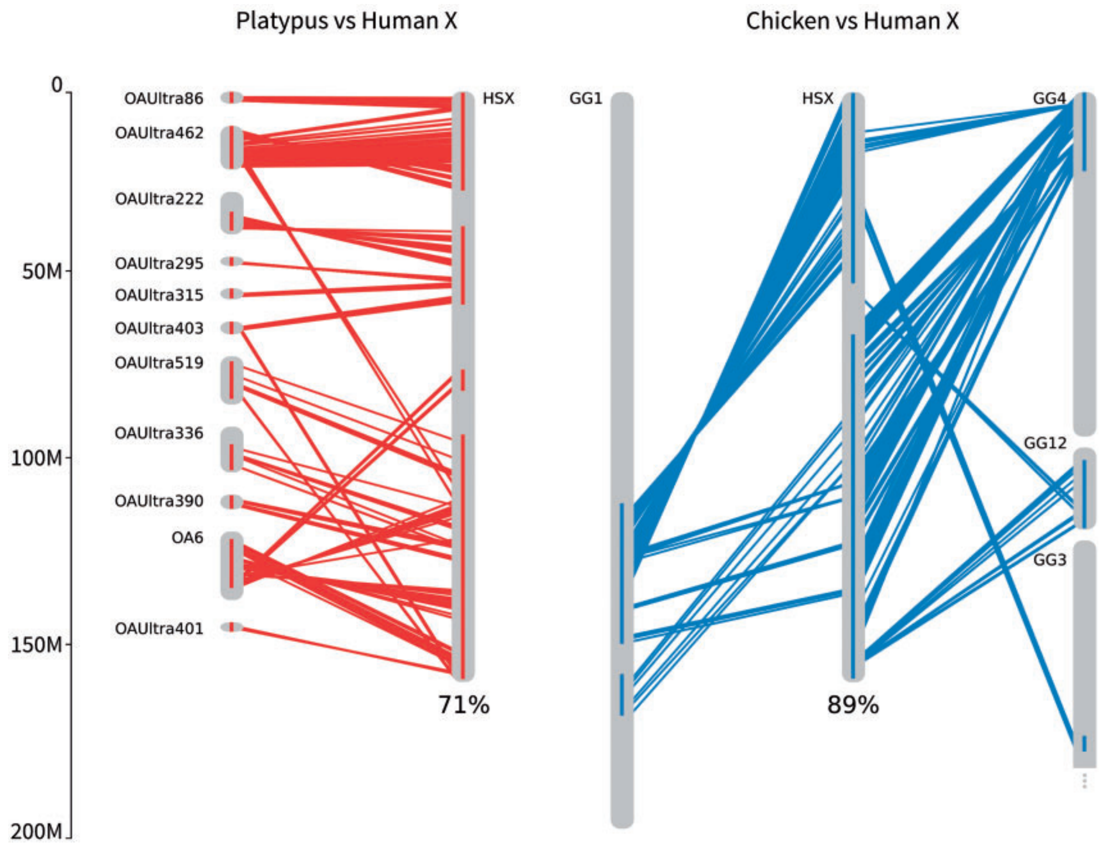


Fig. 1.—Syntenic regions between (*a*) human X (HSX) and platypus chromosome 6 (OA6) and several unmapped contigs (OAUltra) and (*b*) human X (HSX) and chicken chromosomes 1 (GG1), 4 (GG4), 3 (GG3), and 12 (GG12). Lines represent genes in synteny, red for platypus to human, blue for chicken to human. Blocks on chromosomes show regions where single MCScanX alignments are located on the chromosome closer than 10 million base pairs.

519, part of which constitutes the platypus hX-syntenic region. Removing these genes does not qualitatively affect our results (Fisher's exact test, $P > 0.2$).

Between human and chicken, we identified three genes that moved onto the X (from GG8 and two from GG4) and three genes that moved off the X (to HS3 and two to HS2). This is not greater than what we would expect based on the size of the X chromosome (Fisher's exact test, $P > 0.8$, table 1*b*). Again, two of the genes that may have moved onto the X came from regions of GG4 that were close to the hX-syntenic region. These gene movements do not suggest an excess of mt-N gene movement off the human X (table 1*b*, excluding two genes that might not have moved onto the X, Fisher's exact test, $P > 0.3$). One of these genes (ENSP00000362773) was also found to move onto the X in platypus (platypus UltraContig 369 to HSX; chicken GG4 to HSX).

mt-N Gene Abundance on X Syntenic Regions

Our second approach was to compare the abundance of mt-N genes on human X chromosome regions that were syntenic to the identified regions in platypus and chicken. The bias (a measure of mt-N gene density, see Materials and Methods) of mt-N genes does not differ between human X and platypus hX-syntenic regions (Fisher's exact two-tailed test, $P = 0.616$; fig. 2*a*, table 2) or human X and chicken hX-syntenic regions (Fisher's exact two-tailed test, $P = 0.793$; fig. 2*a*, table 2), suggesting that the cause of the underrepresentation on the human X is more likely the result of a random underrepresentation of mt-N genes on the chromosomal regions that formed the human X, rather than intergenomic conflict driving genes off of the X after its formation. We also calculated mt-N gene abundances using species-specific Gene Ontology annotation (GO:0005739) in Biomart to identify mt-N genes. The two approaches to infer mt-N gene function largely agree (platypus 76% overlap; chicken 82% overlap), hence calculating mt-N abundance using Biomart gave qualitatively similar results (table 2, fig. 2*b*, human X and platypus hX-syntenic region, Fisher's exact test, $P = 0.719$; human X and chicken hX-syntenic regions, Fisher's exact test, $P = 0.893$).

Gene Annotation and mt-N Abundance

The measure of abundance (bias) relies on the total number of mt-N genes and total number of genes annotated in each

Table 1
Movement of mt-N Genes On and Off the X between (*a*) Platypus and Human and (*b*) Chicken and Human

Movement	Observed	Expected[a]
(*a*) Platypus → human		
X → A	0	2
A → X	4	4
A → A	132	130
	$P = 0.640$	
(*b*) Chicken → human		
X → A	3	4
A → X	3	4
A → A	92	90
	$P = 0.845$	

Note.—X → A is hX-syntenic to autosome; A → X is autosome to human X syntenic region; A → A is autosome to autosome. *P* value is from Fisher's exact test.

[a] Calculated based on relative size and content of the X chromosome (Betrán et al. 2002; Vibranovski et al. 2009; Toups et al. 2011).

species. This means that measures of bias are susceptible to variation in the quality of genome annotation. The underrepresentation of mt-N genes on the whole of the human X in this study is 0.86 ± 0.22 (bias ± 95% CI), which is less pronounced than the underrepresentation previously reported for the human X chromosome (Drown et al. 2012; Dean et al. 2014). The human genome assembly version has recently been updated from GrCH37 to GrCH38, resulting in changes to the total number of genes and number of mt-N genes, which can account for the different mt-N bias on the human X (bias ± 95% CI, 0.76 ± 0.21 using GrCH37). Gene annotation quality also likely accounts for the overabundance of mt-N genes on the platypus hX-syntenic regions (29 observed mt-N genes and 25 expected), despite a lack of mt-N gene movement off of the X chromosome following X chromosome formation.

mt-N Gene Abundance across Independent X Chromosomes

Across the seven independent X chromosomes studied to date, two (therian mammals and *C. elegans*) show a significant underrepresentation of mt-N genes, three (*Rumex*, platypus and stickleback) exhibit a nonsignificant underrepresentation, and two (*Tribolium* and *Drosophila*) show a nonsignificant overrepresentation (Drown et al. 2012; Dean et al. 2014; Hough et al. 2014). This does not represent a significant overall underrepresentation of mt-N genes on X chromosomes (two-tailed sign-test; 5 of 7, $P = 0.453$). If the distribution of mt-N genes on X chromosomes is explained by variation in ancestral autosomes, we would expect both under- and overrepresentations of mt-N genes on X chromosomes. This is consistent with what we find; however, our ability to detect a significant widespread underrepresentation (i.e., the signature of conflict) is not particularly powerful, with only seven different X chromosomes having been quantified so far. An alternative explanation is that mt-N interactions predispose chromosomes depauperate of mt-N genes to become sex chromosomes, although this predisposition might be rather weak and highly dependent upon the location of genes involved in sex determination.

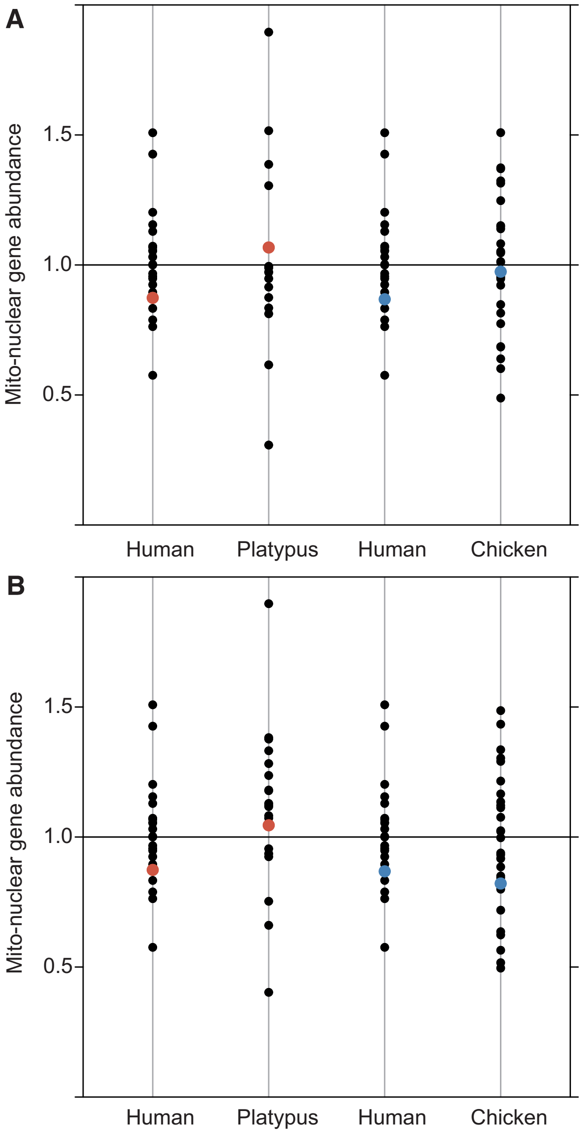


Fig. 2.—Bias of mt-N genes in human, platypus, and chicken. Autosomes in black and hX-syntenic regions with platypus in red, hX-syntenic regions with chicken in blue. (*a*) mt-N genes are inferred using orthology with human mt-N genes, and total gene counts include only those genes that are orthologous between human and platypus or human and chicken. (*b*) mt-N genes are inferred through species-specific annotations in Biomart and gene counts are all annotated genes.

Table 2

Number of mt-N, Total Number of Genes, and the Bias in Distribution of mt-N Genes on the Human X and X-Syntenic Regions Using Gene Orthology to Identify mt-N Genes and Using Species-Specific mt-N Gene Annotations in Biomart

Species	mt-N Genes	Total Genes	Bias	95% CI
Human X	55	820	0.85	0.64–1.06
Platypus hX-syntenic (orthology)	29	309	1.07	0.70–1.43
Platypus hX-syntenic (biomart)	23	381	1.05	0.63–1.45
Human X (syntenic platypus)	46	667	0.87	0.63–1.12
Chicken hX-syntenic (orthology)	64	727	0.97	0.75–1.20
Chicken hX-syntenic (biomart)	52	908	0.83	0.60–1.04
Human X (syntenic chicken)	49	715	0.87	0.64–1.10

Note.—Gene counts are for the hX-syntenic blocks, the boundaries of which are created by merging alignments when alignments were closer than 10 million base pairs. The greater number of orthologous genes on chicken hX-syntenic than on the human X syntenic with chicken region is a consequence of these merged alignments.

Conclusion

Our results suggest that the underrepresentation of mt-N genes on the therian X is not a result of gene movement off of the X chromosome. Rather, the paucity of mt-N genes on the therian X predates the formation of the therian sex chromosomes, and selection has acted mainly to maintain this ancestral distribution after sex chromosome formation. Even though we find no support for conflict driving mt-N genes off the therian X chromosome, random genomic biases in mt-N gene distributions could have important consequences for mt-N coadaptation and potentially for sex chromosome formation. A paucity of mt-N genes on the therian X chromosome means that genes that interact with the mitochondria are less often cotransmitted compared with mt-N genes on autosomes. This might affect rates of coevolution between mitochondria and nuclear genes (e.g., Hill 2014), with possible fitness consequences (Montooth et al. 2010; Meiklejohn et al. 2013).

Materials and Methods

Identification of Ancestral Chromosomes to the Human X Chromosome through Whole-Genome Synteny Analysis

In the first step, we obtained the human (*Homo sapiens*), platypus (*Ornithorhynchus anatinus*), and chicken (*Gallus gallus*) proteomes from Ensembl version 76 (Flicek et al. 2014). We used the longest isoforms as input for BLASTP (Altschul et al. 1990) to detect homologs between the human proteome and both platypus (supplementary table S1, Supplementary Material online) and chicken (supplementary table S2, Supplementary Material online) (*e* value < 10^{-10}). We then used the BLASTP output and positional information as input for MCScanX (Wang et al. 2012), used with default values, to detect homologous chromosomal regions between human and platypus (supplementary table S3, Supplementary Material online) and human and chicken (supplementary table S4, Supplementary Material online). Only genes that have been mapped to a chromosome were included for human and chicken; genes on UltraContigs were included for platypus, as a larger proportion of this genome assembly is currently mapped to scaffolds and contigs rather than chromosomes. The homologous chromosomal regions of the human X chromosome on platypus and chicken chromosomes were identified as the ancestral chromosomes to the human X chromosome. If the individual MCScanX alignments were closer than 10 million base pairs, we merged the alignments into a larger syntenic region to reflect the process of chromosome rearrangement (Burt et al. 1999; Coghlan et al. 2005) and sex chromosome formation (Lahn and Page 1999).

Identification of mt-N Gene Movement

Mt-N genes were identified in human using Gene Ontology annotation (GO:0005739) in Biomart Ensembl Genes 76. To track movement of mt-N genes on and off the X we identified one-to-one orthologs of the 1,572 human mt-N genes in platypus and chicken using reciprocal best hit BLAST (rBBH), with a minimum *e* value of 10^{-10}. Significant hits were ordered by bitscore and a rBBH was only counted when the tophit had a sequence identity larger than 30%. This resulted in 1,064 rBBH between human and platypus, and 1,116 between human and chicken. Of those, 575 rBBH between human and platypus, and 1,087 between human and chicken, were on a sufficiently large scaffold to infer synteny (i.e., Ultra contigs in platypus and chromosomes in chicken).

To identify whether movement of mt-N genes on and off of the X chromosome represents an excess of gene movement, we calculated the expected number of movements based upon the number of genes on source chromosomes and the number of base pairs on the target chromosomes (Betrán et al. 2002; Vibranovski et al. 2009; Toups et al. 2011). Fisher's exact two-tailed tests were used to test whether observed movements were different from expected.

mt-N Abundance

Gene counts of protein-coding genes were calculated using Biomart Ensembl Genes 76. When comparing the abundance of mt-N genes on ancestral X and therian X between species, we used only the regions of the human X chromosome that were identified as syntenic in the other species. The bias of the distribution of mt-N genes on the human X and the platypus and chicken X-syntenic regions was calculated as: Bias = number of mt-N genes/expected number of mt-N genes, where the expected number was calculated as: Expected number = (number of genes in region/total genes) × total mt-N genes.

Mt-N genes in platypus and chicken were identified using two approaches, first, using the orthologous genes to the mt-N genes in human and second, using species-specific Gene Ontology annotation (GO:0005739) in Biomart Ensemble Genes 76. In chicken and platypus GO:0005739 genes are inferred from electronic annotation (evidence code IEA), which includes sequence similarity, database records, and keyword mapping files. As such, the orthology approach and the Biomart approach to infer gene function largely agree, with 76% overlap between the two approaches for platypus and 82% overlap for chicken.

Confidence intervals were calculated using 10,000 bootstrapped samples by randomly sampling genes with replacement and calculating the bias for each iteration. Differences between the expected and actual number of mt-N genes on the human X and platypus or chicken X-syntenic regions were calculated using a Fisher's exact test. Analyses were conducted in R v2.15.1 (R Development Core Team, 2013)

Supplementary Material

Supplementary tables S1–S4 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This research was supported by the European Research Council (grant agreement 260233). The authors thank N. Bloch, P. Harrison, S. Montgomery, A. Wright, and two anonymous reviewers for helpful comments that greatly improved the manuscript.

Literature Cited

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

Arunkumar KP, Mita K, Nagaraju J. 2009. The silkworm Z chromosome is enriched in testis-specific genes. Genetics 182:493–501.

Band MR, et al. 2000. An ordered comparative map of the cattle and human genomes. Genome Res. 10:1359–1368.

Betrán E, Thornton K, Long M. 2002. Retroposed new genes out of the X in *Drosophila*. Genome Res. 12:1854–1859.

Burt DW, et al. 1999. The dynamics of chromosome evolution in birds and mammals. Nature 402:411–413.

Coghlan A, Eichler EE, Oliver SG, Paterson AH, Stein L. 2005. Chromosome evolution in eukaryotes: a multi-kingdom perspective. Trends Genet. 21(12):673–682.

Cortez D, et al. 2014. Origins and functional evolution of Y chromosomes across mammals. Nature 508:488–493.

Dean R, Mank JE. 2014. The role of sex chromosomes in sexual dimorphism: discordance between molecular and phenotypic data. J Evol Biol. 27:1443–1453.

Dean R, Zimmer F, Mank JE. 2014. The potential role of sexual conflict and sexual selection in shaping the genomic distribution of mito-nuclear genes. Genome Biol Evol. 6:1096–1104.

Dehal P, Boore JL. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol. 3:e314.

Drown DM, Preuss KM, Wade MJ. 2012. Evidence of a paucity of genes that interact with the mitochondrion on the X in mammals. Genome Biol Evol. 4:763–768.

Dyall SD, Brown MT, Johnson PJ. 2004. Ancient invasions: from endosymbionts to organelles. Science 304:253–257.

Flicek P, et al. 2014. Ensembl 2014. Nucleic Acids Res. 42(Database issue): D749–D755.

Graves JA. 2006. Sex chromosome specialization and degeneration in mammals. Cell 124:901–914.

Hill GE. 2014. Co-transmission, co-evolution, and conflict sex linkage of nuclear-encoded mitochondrial genes. Heredity 112:469–470.

Hill GE, Johnson JD. 2013. The mitonuclear compatibility hypothesis of sexual selection. Proc Biol Sci. 280:20131314.

Hough J, Agren JA, Barrett SCH, Wright SI. 2014. Chromosomal distribution of cyto-nuclear genes in a dioecious plant with sex chromosomes. Genome Biol Evol. 6:2439–2443.

Innocenti P, Morrow EH, Dowling DK. 2011. Experimental evidence supports a sex-specific selective sieve in mitochondrial genome evolution. Science 332:845–848.

Lahn BT, Page DC. 1999. Four evolutionary strata on the human X chromosome. Science 286:964–967.

Lipinski KJ, et al. 2011. High spontaneous rate of gene duplication in *Caenorhabditis elegans*. Curr Biol. 21:306–310.

Meiklejohn CD, et al. 2013. An incompatibility between a mitochondrial tRNA and its nuclear-encoded tRNA synthetase compromises development and fitness in *Drosophila*. PLoS Genet. 9:e1003238.

Montooth KL, Meiklejohn CD, Abt DN, Rand DM. 2010. Mitochondrial-nuclear epistasis affects fitness within species but does not contribute to fixed incompatibilities between species of *Drosophila*. Evolution 64: 3364–3379.

Murphy WJ, Sun S, Chen ZQ, Pecon-Slattery J, O'Brien SJ. 1999. Extensive conservation of sex chromosome organization between cat and human revealed by parallel radiation hybrid mapping. Genome Res. 9:1223–1230.

Ohno S. 1967. Sex chromosomes and sex-linked genes. Berlin (Germany): Springer.

Partridge L, Hurst LD. 1998. Sex and conflict. Science 281:2003–2008.

R Development Core Team. 2013. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.

Rand DM, Clark AG, Kann LM. 2001. Sexually antagonistic cytonuclear fitness interactions in Drosophila melanogaster. Genetics 159: 173–187.

Raudsepp T, et al. 2004. Exceptional conservation of horse-human gene order on X chromosome revealed by high-resolution radiation hybrid mapping. Proc Natl Acad Sci U S A. 101:2386–2391.

Rogell B, Dean R, Lemos B, Dowling DK. 2014. Mito-nuclear interactions as drivers of gene movement on and off the X-chromosome. BMC Genomics 15:330.

Timmis JN, Aycliffe MA, Huang CY, Martin W. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat Rev Genet. 5:123–135.

Toups MA, Pease JB, Hahn MW. 2011. No excess gene movement is detected off the avian or lepidopteran Z chromosomes. Genome Biol Evol. 3:1463–1472.

Vibranovski MD, Zhang Y, Long M. 2009. General gene movement off the X chromosome in the *Drosophila* genus. Genome Res. 19:897–903.

Wang Y, et al. 2012. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 40:e49.

Wolff JN, Nafisinia M, Sutovsky P, Ballard JWO. 2013. Paternal transmission of mitochondrial DNA as an integral part of mitochondrial inheritance in metapopulations of *Drosophila simulans*. 110 57–62.

Wright AE, Moghadam HK, Mank JE. 2012. Trade-off between selection for dosage compensation and masculinization of the avian Z chromosome. Genetics 192:1433–1445.

Associate editor: Esther Betran

The following article was first published in

Genome Biology and Evolution

Zimmer F, Montgomery SH 2015. Phylogenetic analysis supports a link between DUF1220 domain number and primate brain expansion. *Genome Biology and Evolution* 7(8): 2083-2088, **doi**: 10.1093/gbe/evv122

Genome Biology and Evolution

SMBE

Phylogenetic Analysis Supports a Link between DUF1220 Domain Number and Primate Brain Expansion

Fabian Zimmer and Stephen H. Montgomery*

Department of Genetics, Evolution & Environment, University College London, United Kingdom

*Corresponding author: E-mail: stephen.montgomery@cantab.net.

Accepted: June 19, 2015

Abstract

The expansion of DUF1220 domain copy number during human evolution is a dramatic example of rapid and repeated domain duplication. Although patterns of expression, homology, and disease associations suggest a role in cortical development, this hypothesis has not been robustly tested using phylogenetic methods. Here, we estimate DUF1220 domain counts across 12 primate genomes using a nucleotide Hidden Markov Model. We then test a series of hypotheses designed to examine the potential evolutionary significance of DUF1220 copy number expansion. Our results suggest a robust association with brain size, and more specifically neocortex volume. In contradiction to previous hypotheses, we find a strong association with postnatal brain development but not with prenatal brain development. Our results provide further evidence of a conserved association between specific loci and brain size across primates, suggesting that human brain evolution may have occurred through a continuation of existing processes.

Key words: autistic spectrum disorder, brain evolution, DUF1220 domains, NBPF, primates.

Introduction

The molecular targets of selection favoring brain expansion during human evolution have been sought by identifying dramatic, lineage-specific shifts in evolutionary rate. The increase in DUF1220 domains during human evolution provides one of the most dramatic increases in copy number (Popesco et al. 2006; Dumas et al. 2012). A single copy of this protein domain is found in *PDE4DIP* in most mammalian genomes. In primates, this ancestral domain has been duplicated many times over, reaching its peak abundance in humans where several hundred DUF1220 domains exist across 20–30 genes in the Nuclear Blastoma Breakpoint Family (NBPF) (Vandepoele et al. 2005; Dumas et al. 2012). The majority of these map to 1q21.1, a chromosomal region with complex, and unstable genomic architecture (O'Bleness et al. 2012, 2014).

Interspecific DUF1220 counts show a pattern of phylogenetic decay with increasing distance from humans (Popesco et al. 2006; Dumas and Sikela 2009; Dumas et al. 2012). In humans, DUF1220 dosage has also been linked to head circumference (Dumas et al. 2012), and severe neurodevelopmental disorders, including autism spectrum disorder (ASD) and microcephaly (Dumas et al. 2012; Davis et al. 2014). The severity of ASD impairments is also correlated with 1q21.1 DUF1220 copy number suggesting a dosage effect (Davis et al. 2014). Taken together, these observations led to the suggestion that the expansion of DUF1220 copy number played a primary role in human brain evolution (Dumas and Sikela 2009; Keeney, Dumas, et al. 2014).

Although functional data are limited, they provide some indication of how DUF1220 domain copy number count influences brain development. DUF1220 domains are highly expressed during periods of cortical neurogenesis, suggesting a potential role in prolonging the proliferation of neural progenitors by regulating centriole and microtubule dynamics to control key cell fate switches critical for neurogenesis (Keeney, Davis, et al. 2014). *PDE4DIP*, which contains the ancestral DUF1220 domain, does indeed associate with the spindle poles (Popesco et al. 2006) and is homologous to *CDK5RAP2*, a centrosomal protein essential for neural proliferation (Bond et al. 2005; Buchman et al. 2010), which coevolved with brain mass across primates (Montgomery et al. 2011).

Two previous analyses report a significant association between DUF1220 copy number and brain mass, cortical neuron number (Dumas et al. 2012), cortical gray and white matter, surface area, and gyrification (Keeney, Davis, et al. 2014). However, several limitations in these analyses restrict confidence in the results. First, DUF1220 copy number was assessed across species using a BLAT/BLAST (BLAST-like alignment tool/Basic Local Alignment Search Tool) analysis with a query sequence from humans, which introduces a bias that could partly explain the observed phylogenetic decay. Second, counts were not restricted to those domains occurring in functional exonic sequence and therefore many DUF1220 domains found in human pseudogenes were included in the analyses. Third, the analyses were limited to a small number of species (4–8 primates), using parametric statistics that may not be suitable for count data, and which do not correct for phylogenetic nonindependence (Felsenstein 1985). This is not a negligible issue, as it can result in the overestimation of statistical significance (Carvalho et al. 2006). Finally, previous phenotypic associations have been reported for multiple cortical phenotypes all of which are strongly correlated with one another or are nonindependent. Therefore, to date, these studies have not provided evidence for a specific association with neocortex size, neither have they tested the strength of the association with different periods of brain development, which may provide new clues as to the functional relevance of DUF1220 domain copy number.

Here, we use nucleotide Hidden Markov Models (HMMs) (HMMER3; Eddy 2011) to more accurately query the DUF1220 domain number of distantly related genomes. After filtering these counts to limit the analysis to exonic sequence, we use phylogenetic comparative methods that correct for nonindependence to test whether DUF1220 copy number is robustly associated with brain size, whether this is due to an association with pre- or postnatal brain development, and whether the association is specific to the neocortex.

Table 1

DUF1220 Count Data

Species	O'Bleness et al. (2012)	nHMM	
		Whole Genome	Functional Exonic with CM Promoter
Homo sapiens	272	302	262
Pan troglodytes	125	138	32
Gorilla gorilla	99	97	32
Pongo abelii	92	101	27
Nomascus leucogenys	53	59	6
Papio anubis	—	75	15
Chlorocebus sabaeus	—	48	16
Macaca mulatta	35	74	10
Callithrix jacchus	31	75	9
Tarsius syrichta	—	47	2
Microcebus murinus	2	4	1
Otolemur garnettii	3	4	2

Results

We confirm significant interspecific variation in DUF1220 counts across primates (table 1, fig. 1). Phylogenetic Generalized Least Square (PGLS) regressions (Pagel 1999) using square-root, or $\log_{10}$-transformed DUF1220 counts support previous reports of an association with brain volume (SQRT: $t_{10} = 3.165$, $P = 0.005$, $R^2 = 0.455$; $\log_{10}$: $t_{10} = 4.770$, $P < 0.001$, $R^2 = 0.655$). The same associations are also found after excluding *Homo sapiens* from the analysis (SQRT: $t_9 = 3.810$, $P = 0.002$, $R^2 = 0.569$; $\log_{10}$: $t_9 = 3.952$, $P = 0.002$, $R^2 = 0.586$). However, these data transformations may not be appropriate for count data where models based on Poisson distributions provide more accurate results (O'Hara and Kotze 2010).

Using a Bayesian approach that corrects for phylogenetic nonindependence and fits a Poisson distribution to the DUF1220 count data (MCMCglmm; Hadfield 2010), we again find evidence that CM-associated exonic DUF1220

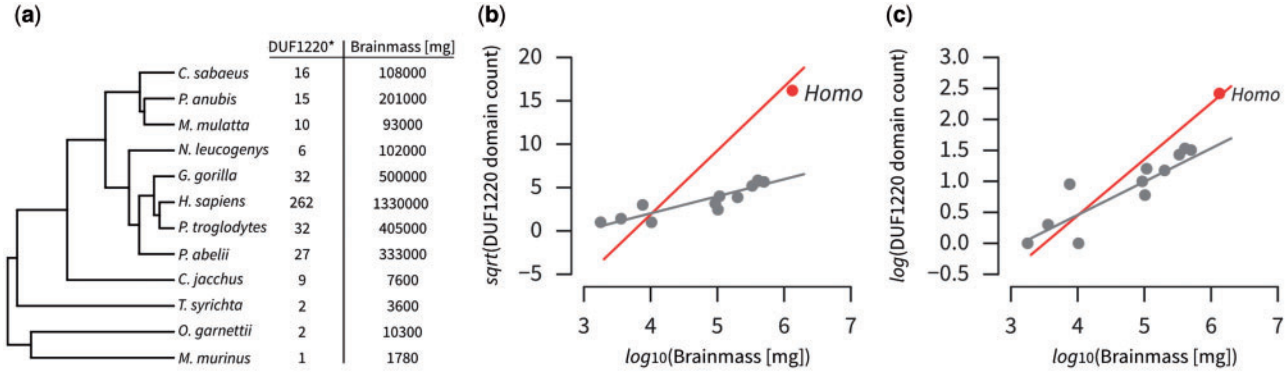


Fig. 1.— (*a*) Phylogeny of Ensembl primate genomes showing the number of DUF1220 domains in functional, annotated genes with a CM promoter, and brain mass. (*b*) The relationship between square-root transformed DUF1220 counts and $\log_{10}$(brain mass), and (c) the relationship between $\log_{10}$ transformed DUF1220 counts and $\log_{10}$(brain mass). The regression lines are shown with (red) and without (gray) the inclusion of the *H. sapiens* data. In all cases, they are significant.

counts are associated with brain mass across primates ($n = 12$, posterior mean = 1.927, 95% confidence interval [CI] = 0.800–3.040, $P_{MCMC} = 0.001$). This association is robust to the exclusion of *H. sapiens* (posterior mean = 1.271, 95% CI = 0.490–2.019, $P_{MCMC} = 0.003$), and found when hominoids ($n = 5$, posterior mean = 3.679, 95% CI = 0.966–6.258, $P_{MCMC} = 0.018$) or anthropoids ($n = 9$, posterior mean = 2.019, 95% CI = 0.352–3.684, $P_{MCMC} = 0.010$) are analyzed alone, suggesting a consistent phylogenetic association. When body mass is included as a cofactor in the model, the positive association is restricted to brain mass (table 2*a*, fig. 1*a*).

Separation of pre- and postnatal development specifically links DUF12220 number to postnatal brain growth. Analyzed separately, the association with prenatal brain growth is weaker ($n = 11$, posterior mean = 1.758, 95% CI = –0.039 to 3.543, $P_{MCMC} = 0.023$) than with postnatal brain growth (posterior mean = 1.839, 95% CI = 0.895–2.808, $P_{MCMC} = 0.001$). If both traits are included in the same model, only the positive association with postnatal brain growth remains (table 2*b*, fig. 2*b* model 1). Multiple regression analysis also confirms that the association is specific to postnatal brain growth, rather than postnatal body growth (table 2*b* model 2).

Table 2

MCMCglmm Results of Multivariate Models

Model	Posterior Mean	95% CI	P_{MCMC}
(*a*) Brain Mass and Body Mass			
1. log(brain mass)	4.105	2.163 to 6.000	0.001
+ log(body mass)	−1.986	−3.544 to −3.900	0.988
(*b*) Prenatal and Postnatal Growth			
1. log(prenatal brain growth)	−2.158	−4.471 to 0.106	0.967
+ log(postnatal brain growth)	3.319	1.470 to 4.982	0.002
2. log(postnatal brain growth)	2.910	1.641 to 4.151	<0.001
+ log(postnatal body growth)	−1.241	−2.442 to −0.052	0.977
(*c*) Brain Regions			
1. log(neocortex volume)	5.961	0.720 to 11.173	0.014
+ log(RoB volume)	−5.817	−13.322 to 1.120	0.953
2. log(cerebellum volume)	3.699	−5.857 to 12.611	0.186
+ log(RoB volume)	−2.435	−13.869 to 10.132	0.681
3. log(neocortex volume)	6.076	−0.139 to 12.5712	0.025
+ log(cerebellum volume)	−0.369	−9.5128 to 8.961	0.526
+ log(RoB volume)	−5.494	−15.814 to 5.288	0.872

Finally, we not only examined the hypothesized relationship with neocortex volume (e.g., Keeny, Davis, et al. 2014; Keeny, Dumas, et al. 2014), but also considered cerebellum volume, as this region coevolves with the neocortex (Barton and Harvey 2000), has expanded in apes (Barton and Venditti 2014), and shows high levels of NBPF expression (Popesco et al. 2006). When the rest-of-the-brain (RoB) is included as a cofactor, to account for variation in overall brain size, a positive association is found for neocortex volume but not cerebellum volume (table 2*c* models 1-3, fig. 2*c*).

Discussion

Our phylogenetic analyses substantiate the hypothesis that the increase in DUF1220 number coevolves with brain mass (Dumas et al. 2012; Keeney, Davis, et al. 2014), and may contribute to the proximate basis of primate brain evolution. We extend the results of previous studies by demonstrating specific associations with neocortex volume, and postnatal brain growth rather than prenatal brain growth. Together these results imply a role for DUF1220 in evolutionary changes in the maturation and postnatal development of the neocortex. Previous hypotheses concerning the phenotypic relevance of DUF1220 domain number have focused on their possible contribution to neurogenesis (Dumas and Sikela 2009; Keeny, Davis, et al. 2014; Keeny, Dumas, et al. 2014). This is supported by homology to genes with known functions in cell cycle dynamics (Popesco et al. 2006; Thornton and Woods

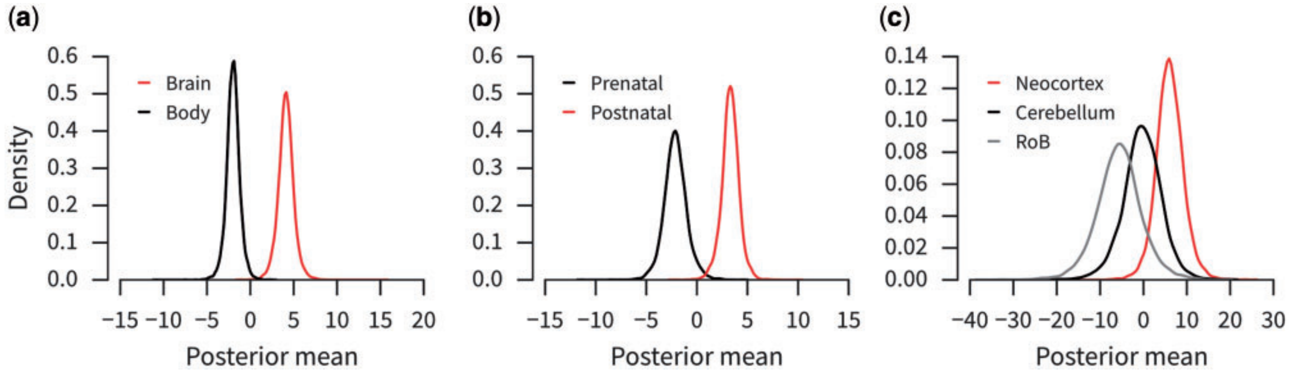


FIG. 2.— (*a*) Posterior means of the association between DUF1220 count and brain mass (red) and body mass (black). (*b*) Posterior means of the association between DUF1220 count and postnatal brain growth (red) and prenatal brain growth (black). (c) Posterior means of the association between DUF1220 count and neocortex volume (red), cerebellum volume (solid black), and rest-of-brain volume (dashed black).

2009), relevant spatial and temporal expression patterns (Keeney, Davis, et al. 2014), and an effect on the proliferation of neuroblastoma cell cultures (Vandepoele et al. 2008). However, a direct effect of variation in DUF1220 domain number on neural proliferation has not been demonstrated (Keeney et al. 2015).

If DUF1220 domains do regulate neurogenesis, we would expect them to coevolve with prenatal brain growth, as cortical neurogenesis is restricted to prenatal development (Bhardwaj et al. 2006). Our results instead suggest a robust and specific relationship with postnatal brain development. Existing data on DUF1220 domain function suggest two potential roles that may explain this association: 1) a contribution to axonogenesis through initiating and stabilizing microtubule growth in dendrites; and 2) a potential role in apoptosis during brain maturation. Both hypotheses are consistent with the reported association between variation in DUF1220 dosage and ASD (Davis et al. 2014). Indeed, an emphasis on postnatal brain growth is potentially more relevant for ASD, which develops postnatally, accompanied by a period of accelerated brain growth in early postnatal development (Courchesne et al. 2011).

Microtubule assembly is essential for dendritic growth and axonogenesis (Conde and Cáceres 2009). *PDE4DIP*, which contains the ancestral DUF1220 domain, has known functions in microtubule nucleation, growth, and cell migration (Roubin et al. 2013). There is also evidence that NBPF1 interacts with a key regulator of Wnt signaling (Vandepoele et al. 2010) that has important roles in neuronal differentiation, dendritic growth, and plasticity (Inestrosa and Varela-Nallar 2014). Consistent with this function, DUF1220 domains are highly expressed in the cell bodies and dendrites of adult neurons (Popesco et al. 2006). A role for DUF1220 domains in synaptogenesis could potentially explain the association with ASD severity (Davis et al. 2014). ASD is associated with abnormalities in cortical minicolumns (Casanova et al. 2002) and cortical white matter (Hazlett et al. 2005; Courchesne et al. 2011), both of which suggest a disruption of normal neuronal maturation (Courchesne and Pierce 2005; Minshew and Williams 2007).

Alternatively, NBPF genes are also known to interact with NF-κB (Zhou et al. 2013), a transcription factor implicated in tumor progression, with a range of roles including apoptosis and inflammation (Karin and Lin 2002; Perkins 2012). Postnatal apoptosis has a significant influence on brain growth (Kuan et al. 2000; Polster et al. 2003; Madden et al. 2007), including regulating neuronal density (Sanno et al. 2010), and apoptotic genes may have been targeted by selection in relation to primate brain expansion (Vallender and Lahn 2006). Disruption of apoptosis causes microcephaly (Poulton et al. 2011), potentially explaining the association between DUF1220 dosage and head circumference (Dumas et al. 2012). The association of NF-κB with inflammatory diseases (Tak et al. 2001) is also intriguing, given the growing evidence that the inflammatory response is linked to the risk and severity of ASD (Meyer et al. 2011; Depino 2012).

If DUF1220 domain number does contribute to the evolution of postnatal brain growth, this contrasts with results of previously studied candidate genes with known roles in neurogenesis that coevolve with prenatal brain growth (Montgomery et al. 2011). This suggests a two-component model of brain evolution where selection targets one set of genes to bring about an increase in neuron number (e.g., Montgomery et al. 2011; Montgomery and Mundy 2012a, 2012b), and an independent set of genes to optimize neurite growth and connectivity (e.g., Charrier et al. 2012). NBPF genes may fall into the latter category. This two-component model is consistent with comparative analyses that indicate pre- and postnatal brain developments evolve independently, and must therefore be relatively free of reciprocal pleiotropic effects (Barton and Capellini 2011).

Finally, these results add further evidence that many of the genetic changes that contribute to human evolution will be based on the continuation or exaggeration of conserved gene-phenotype associations that contribute to primate brain evolution more broadly (Montgomery et al. 2011; Scally et al. 2012). Understanding the commonalities between human and nonhuman primate brain evolution is therefore essential to understand the genetic differences that contribute the derived aspects of human evolution.

Materials and Methods

Counting DUF1220 Domains

HMMER3.1b (Eddy 2011) was used to build an HMM from the DUF1220 (PF06758) seed alignment stored in the PFAM database (Finn et al. 2014). The longest isoforms for all proteomes of 12 primate genomes from Ensembl v.78 (Cunningham et al. 2015) (fig. 1*a*) were searched using the protein DUF1220 HMM (hmmsearch, *E* value < 1e-10) (supplementary table S1, Supplementary Material online). We extracted the corresponding cDNA regions to build a DUF1220 nucleotide profile HMM (nHMM) using a MAFFT sequence alignment, allowing for more sensitive analysis across a broad phylogenetic range. The DUF1220 nHMM was used to search the complete genomic DNA for all 12 species. These counts were filtered to remove any DUF1220 domains not located in annotated exonic sequence, or located in known pseudogenes.

We next filtered our counts to limit them to exonic sequence in close proximity to the NBPF-specific Conserved-Mammal (CM) promoter (O'Bleness et al. 2012). To do so, we built a nucleotide HMM for the CM promoter based on a MAFFT (Katoh et al. 2002) alignment of the 900-bp CM region upstream of human genes *NBPF4*, *NBPF6*, and *NBPF7*. Using this CM promoter nHMM, we searched 1,000-bp up- and downstream of genes containing DUF1220 domains for significant CM promoter hits (nhmmer, *E* value < 1e-10). This provided final counts for DUF1220 domains within exonic regions and associated with the CM promoter (table 1). These counts were used in subsequent phylogenetic analyses. In the supplementary information, Supplementary Material online, we compare our counts with previous estimates and discuss possible sources of error. All scripts and data used in the analysis are freely available from: https://github.com/qfma/duf1220

Phylogenetic Gene-Phenotype Analysis

PGLS regressions were performed using log-transformed phenotypic data and log- or square root-transformed DUF1220 count data in BayesTraits (Pagel 1999). Phylogenetic multivariate generalized mixed models were implemented using a Bayesian approach in MCMCglmm (Hadfield 2010), to test for phylogenetically corrected associations between DUF1220 counts and log-transformed phenotypic data (supplementary table S2, Supplementary Material online). All analyses were performed using a Poisson distribution, as recommended for count data (O'Hara and Kotze 2010), with uninformative, parameter expanded priors for the random effect (G: V = 1,n ν = 1, alpha.ν = 0, alpha.V = 1,000; R: V = 1, ν = 0.002) and default priors for the fixed effects. Phylogenetic relationships were taken from the 10k Trees project (Arnold et al. 2010). We report the posterior mean of the cofactor included in each model and its 95% CIs, and the probability that the parameter value is greater than 0 (P_{MCMC}) as we specifically hypothesize a positive association (Dumas et al. 2012). Alternative data treatments lead to similar conclusions (supplementary information, Supplementary Material online).

Acknowledgments

The authors thank James Sikela, Majesta O'Bleness, Chris Venditti, Charlotte Montgomery, Andrew Moore, and Jarrod Hadfield for advice and comments, and Judith Mank's lab at UCL for support. S.H.M. thanks the Leverhulme Trust for funding.

Supplementary Material

Supplementary information, figures S1–S3, and tables S1–S3 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Literature Cited

Arnold C, Matthews LJ, Nunn CL. 2010. The 10kTrees website: a new online resource for primate phylogeny. Evol Anthropol. 19:114–118.

Barton RA Capellini I. 2011. Maternal investment, life histories, and the costs of brain growth in mammals. Proc Natl Acad Sci U S A. 108:6169–6174.

Barton RA, Harvey PH. 2000. Mosaic evolution of brain structure in mammals. Nature 405:1055–1058.

Barton RA, Venditti C. 2014. Report rapid evolution of the cerebellum in humans and other Great Apes. Curr Biol. 24:2440–2444.

Bhardwaj RD, et al. 2006. Neocortical neurogenesis in humans is restricted to development. Proc Natl Acad Sci U S A. 103:12564–12568.

Bond J, et al. 2005. A centrosomal mechanism involving CDK5RAP2 and CENPJ controls brain size. Nat Genet. 37:353–355.

Buchman JJ, et al. 2010. Cdk5rap2 interacts with pericentrin to maintain the neural progenitor pool in the developing neocortex. Neuron 66:386–402.

Carvalho P, Felizola Diniz-Filho JAF, Bini LM. 2006. Factors influencing changes in trait correlations across species after using phylogenetic independent contrasts. Evol Ecol. 20(6):591-602.

Casanova MF, Buxhoeveden DP, Cohen M, Switala AE, Roy EL. 2002. Minicolumnar pathology in dyslexia. Ann Neurol. 52:108–110.

Charrier C, et al. 2012. Inhibition of SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation. Cell 149:923–935.

Conde C, Cáceres A. 2009. Microtubule assembly, organization and dynamics in axons and dendrites. Nat Rev Neurosci. 10:319–332.

Courchesne E, et al. 2011. Unusual brain growth patterns in early life in patients with autistic disorder: an MRI study. Neurology 76:2111.

Courchesne E, Pierce K. 2005. Why the frontal cortex in autism might be talking only to itself: local over-connectivity but long-distance disconnection. Curr Opin Neurobiol. 15:225–230.

Cunningham F, et al. 2015. Ensembl 2015. Nucleic Acids Res. 43:D662–D669.

Davis JM, et al. 2014. DUF1220 dosage is linearly associated with increasing severity of the three primary symptoms of autism. PLoS Genet. 10:1–5.

Depino AM. 2012. Peripheral and central inflammation in autism spectrum disorders. Mol Cell Neurosci. 53:69–76.

Dumas L, Sikela JM. 2009. DUF1220 domains, cognitive disease, and human brain evolution. Cold Spring Harb Symp Quant Biol. 74:375–382.

Dumas LJ, et al. 2012. DUF1220-domain copy number implicated in human brain-size pathology and evolution. Am J Hum Genet. 91:444–454.

Eddy SR. 2011. Accelerated profile HMM searches. PLoS Comput Biol. 7(10):e1002195.

Felsenstein J. 1985. Phylogenies and the comparative method. Am Nat. 125:1–15.

Hadfield JD. 2010. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. J Stat Softw. 33:1–22.

Hazlett HC, et al. 2005. Magnetic resonance imaging and head circumference study of brainsize in autism: birth through age 2 years. Arch Gen Psychiatry. 62:1366–1376.

Inestrosa NC, Varela-Nallar L. 2014. Wnt signalling in neuronal differentiation and development. Cell Tissue Res. 359:215–223.

Karin M, Lin A. 2002. NF-κB at the crossroads of life and death. Nat Immunol. 3:221–227.

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

Keeney J, Dumas L, Sikela J. 2014. The case for DUF1220 Domain dosage as a primary contributor to anthropoid brain expansion Front Hum Neurosci. 8:1–11.

Keeney JG, Davis JM, et al. 2014. DUF1220 protein domains drive proliferation in human neural stem cells and are associated with increased cortical volume in anthropoid primates. Brain Struct Funct. 1–8.

Keeney JG, et al. 2015. Generation of mice lacking DUF1220 protein domains: effects on fecundity and hyperactivity. Mamm Genome. 26:33–42.

Kuan CY, Roth KA, Flavell RA, Rakic P. 2000. Mechanisms of programmed cell death in the developing brain. Trends Neurosci. 23:291–297.

Madden SD, Donovan M, Cotter TG. 2007. Key apoptosis regulating proteins are down-regulated during postnatal tissue development. Int J Dev Biol. 51:415–425.

Meyer U, Feldon J, Dammann O. 2011. Schizophrenia and autism: both shared and disorder-specific pathogenesis via perinatal inflammation? Pediatr Res 69:26–33.

Minshew NJ, Williams DL. 2007. The new neurobiology of autism. Arch Neurol. 64:945–950.

Montgomery SH, Capellini I, Venditti C, Barton RA, Mundy NI. 2011. Adaptive evolution of four microcephaly genes and the evolution of brain size in anthropoid primates. Mol Biol Evol. 28:625–638.

Montgomery SH, Mundy NI. 2012a. Evolution of ASPM is associated with both increases and decreases in brain size in primates. Evolution 66:927–932.

Montgomery SH, Mundy NI. 2012b. Positive selection on NIN, a gene involved in neurogenesis, and primate brain evolution. Genes Brain Behav. 11:903–910.

O'Bleness MS, et al. 2012. Evolutionary history and genome organization of DUF1220 protein domains. G3 (Bethesda) 2:977–986.

O'Bleness M, et al. 2014. Finished sequence and assembly of the DUF1220-rich 1q21 region using a haploid human genome. BMC Genomics 15:387.

O'Hara RB, Kotze DJ. 2010. Do not log-transform count data. Methods Ecol Evol. 1:118–122.

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

Perkins ND. 2012. The diverse and complex roles of NF-κB subunits in cancer. Nat Rev Cancer. 12(2):121–132.

Polster BM, Robertson CL, Bucci CJ, Suzuki M, Fiskum G. 2003. Postnatal brain development and neural cell differentiation modulate mitochondrial Bax and BH3 peptide-induced cytochrome c release. Cell Death Differ. 10:365–370.

Popesco MC, et al. 2006. Human lineage-specific amplification, selection, and neuronal expression of DUF1220 domains. Science 313:1304–1307.

Poulton CJ, et al. 2011. Microcephaly with simplified gyration, epilepsy, and infantile diabetes linked to inappropriate apoptosis of neural progenitors. Am J Hum Genet. 89:265–276.

Roubin R, et al. 2013. Myomegalin is necessary for the formation of centrosomal and Golgi-derived microtubules. Biol Open. 2:238–250.

Sanno H, et al. 2010. Control of postnatal apoptosis in the neocortex by RhoA-subfamily GTPases determines neuronal density. J Neurosci. 30:4221–4231.

Scally A, et al. 2012. Insights into hominid evolution from the gorilla genome sequence. Nature 483:169–175.

Tak PP, Firestein GS, Tak PP, Firestein GS. 2001. NF-κB: a key role in inflammatory diseases. J Clin Invest.107:7–11.

Thornton GK, Woods CG. 2009. Primary microcephaly: do all roads lead to Rome? Trends Genet 25:501–510.

Vallender EJ, Lahn BT. 2006. A primate-specific acceleration in the evolution of the caspase-dependent apoptosis pathway. Hum Mol Genet. 15:3034–3040.

Vandepoele K, et al. 2008. A constitutional translocation t(1;17)(p36.2;q11.2) in a neuroblastoma patient disrupts the human NBPF1 and ACCN1 genes. PLoS One 3(5):e2207.

Vandepoele K, Staes K, Andries V, van Roy F. 2010. Chibby interacts with NBPF1 and clusterin, two candidate tumor suppressors linked to neuroblastoma. Exp Cell Res. 316:1225–1233.

Vandepoele K, Van Roy N, Staes K, Speleman F, Van Roy F. 2005. A novel gene family NBPF: intricate structure generated by gene duplications during primate evolution. Mol Biol Evol. 22:2265–2274.

Zhou F, et al. 2013. NBPF is a potential DNA-binding transcription factor that is directly regulated by NF-κB. Int J Biochem Cell Biol. 45:2479–2490.

Associate editor: George Zhang

The following article was first published in

Proceedings of the National Academy of Sciences, USA

Harrison PW, Wright AE, **Zimmer F**, Dean R, Montgomery SH, Pointer MA, Mank JE (2015) Sexual selection drives evolution and rapid turnover of male gene expression. *Proceedings of the National Academy of Sciences, USA* 112: 4393-4398 Copyright (2015) National Academy of Sciences.
doi: 10.1073/pnas.1501339112

The following article was first published in

Molecular Biology and Evolution

Dean R, Harrison PW, Wright AE, **Zimmer F**, Mank JE (2015) Positive selection underlies Faster-Z evolution of gene expression in birds. *Molecular Biology and Evolution* 32: 2646-2656
doi: 10.1093/molbev/msv138

Positive Selection Underlies Faster-Z Evolution of Gene Expression in Birds

Rebecca Dean,*[1] Peter W. Harrison,[1] Alison E. Wright,[1] Fabian Zimmer,[1] and Judith E. Mank[1]

[1]Department of Genetics, Evolution and Environment, University College London, London, United Kingdom

***Corresponding author:** E-mail: r.dean@ucl.ac.uk.

Associate editor: Doris Bachtrog

Abstract

The elevated rate of evolution for genes on sex chromosomes compared with autosomes (Fast-X or Fast-Z evolution) can result either from positive selection in the heterogametic sex or from nonadaptive consequences of reduced relative effective population size. Recent work in birds suggests that Fast-Z of coding sequence is primarily due to relaxed purifying selection resulting from reduced relative effective population size. However, gene sequence and gene expression are often subject to distinct evolutionary pressures; therefore, we tested for Fast-Z in gene expression using next-generation RNA-sequencing data from multiple avian species. Similar to studies of Fast-Z in coding sequence, we recover clear signatures of Fast-Z in gene expression; however, in contrast to coding sequence, our data indicate that Fast-Z in expression is due to positive selection acting primarily in females. In the soma, where gene expression is highly correlated between the sexes, we detected Fast-Z in both sexes, although at a higher rate in females, suggesting that many positively selected expression changes in females are also expressed in males. In the gonad, where intersexual correlations in expression are much lower, we detected Fast-Z for female gene expression, but crucially, not males. This suggests that a large amount of expression variation is sex-specific in its effects within the gonad. Taken together, our results indicate that Fast-Z evolution of gene expression is the product of positive selection acting on recessive beneficial alleles in the heterogametic sex. More broadly, our analysis suggests that the adaptive potential of Z chromosome gene expression may be much greater than that of gene sequence, results which have important implications for the role of sex chromosomes in speciation and sexual selection.

Key words: **female heterogamety, gene expression divergence, selection, drift, Fast-X, sex chromosomes.**

Introduction

The unique properties of the sex chromosomes are thought to influence rates of evolution for the genes they contain, and comparisons between the sex chromosomes and autosomes are important for understanding the role that dominance, effective population size and recombination play in adaptive evolution. For both X and Z chromosomes, hemizygosity and lower relative effective population size (N_E) of sex chromosomes can lead to an increased rate of functional change in comparison to autosomes (Charlesworth et al. 1987; Vicoso and Charlesworth 2006), a process termed Fast-X or Fast-Z evolution.

In female heterogametic sex chromosome systems, the single copy of the Z chromosome in females means that recessive beneficial alleles are always exposed to selection when expressed in this sex, leading to greater rates of fixation of recessive advantageous alleles. This would result in the Fast-Z effect due to adaptive evolution. Alternatively, Fast-Z can occur as a result of the reduced N_E of the Z compared with the autosomes. When male and female reproductive success are equal, there are only three Z chromosomes for every four autosomes ($N_{EZ} = 3/4\ N_{EA}$). The reduction in N_{EZ} leads to a reduction in the efficacy of purifying selection on the Z chromosome (Caballero 1995; Laporte and Charlesworth 2002) and drift has greater potential to fix mildly deleterious alleles (Vicoso and Charlesworth 2009). Differentiating the role of hemizygosity versus reduced N_E in rates of evolution for sex chromosomes is essential for determining the relative role of adaptive evolution versus genetic drift in sex chromosome evolution, with important implications for sexual selection and speciation (e.g., Haldane 1922; Kirkpatrick and Hall 2004).

Fast-Z evolution has been broadly detected in studies of coding sequence in birds (Mank et al. 2007; Mank, Nam, et al. 2010; Corl and Ellegren 2012; Wright et al. 2015), snakes (Vicoso et al. 2013), and moths (Sackton et al. 2014). Most examples of Fast-Z sequence evolution have mainly been attributed to drift (Mank, Nam, et al. 2010; Vicoso et al. 2013; Wright et al. 2015) although evidence from silk moths suggests positive selection (Sackton et al. 2014). Moreover, drift may be particularly strong on Z chromosomes due to sexual selection. Increasing variance in male reproductive success, such as that produced by sexual selection (Wade 1979; Andersson 1994), reduces relative N_{EZ} below $3/4\ N_{EA}$, unlike male heterogametic systems (Mank, Vicoso, et al. 2010). Recent estimates of N_{EZ} in birds have been significantly less than $3/4\ N_{EA}$ (Corl and Ellegren 2012; Wang et al. 2014; Wright et al. 2015) potentially resulting in elevated levels of genetic drift for Z-linked genes.

Studies of Fast-Z evolution have so far focused on coding sequence data of orthologous genes to compare rates of change on the Z chromosome versus the autosomes (Mank et al. 2007; Mank, Nam, et al. 2010; Corl and Ellegren 2012;

Vicoso et al. 2013; Sackton et al. 2014; Wright et al. 2015). Genes that are orthologous across species tend to be under high purifying selection (Wang et al. 2007) and as such this may limit the ability of gene sequence studies to detect adaptive signals of Fast-Z. Although gene expression studies also use orthologous genes, sequence and expression can show different patterns of evolution, even for the same locus. For example, purifying selection may act more weakly on expression of conserved orthologous genes if the regions regulating gene expression are less conserved, thus allowing greater capacity for adaptive evolution of gene expression. Additionally, gene expression evolution may also be influenced by *trans*-regulation from different chromosomes (Meisel et al. 2012; Meisel and Connallon 2013; Meiklejohn et al. 2014). Studies of expression evolution on sex chromosomes may therefore be particularly informative for understanding the nature of gene expression evolution (Kayserili et al. 2012; Meisel et al. 2012; Meisel and Connallon 2013), and for identifying the adaptive potential of sequence versus expression evolution (Stern and Orgogozo 2008).

In order to perform the first test of Fast-Z evolution of global gene expression, we built de novo transcriptome assemblies from somatic and gonadal tissue from captive males and females of six species of the Galloanserae, including turkey (*Meleagris gallopavo*), pheasant (*Phasianus colchicus*), peafowl (*Pavo cristatus*), guinea fowl (*Numida meleagris*), swan goose (*Anser cygnoides*), and mallard duck (*Anas platyrhynchos*) (Harrison et al. 2015; Wright et al. 2015). Our data indicate that gene expression on the Z chromosome evolves more rapidly than that on the autosomes, consistent with previous studies of Fast-Z in coding sequence. However, we observe more pronounced Fast-Z in females than males, suggesting that unlike protein coding sequence, Fast-Z in avian gene expression is primarily adaptive in nature. Together, our results suggest that gene expression on the Z chromosome may have a greater adaptive potential than coding sequence, a finding with important implications for sexual selection and speciation.

Results

Faster-Z Evolution of Gene Expression

We calculated the pairwise similarity in expression separately for each sex, using Spearman's rho correlation coefficient (ρ) (Brawand et al. 2011; Meisel et al. 2012). We used pheasant as the reference point (i.e., comparing expression of each of the other five species to pheasant) in order to achieve even phylogenetic spacing of taxa, which maximizes our power to test for differences in the slope of ρ between the Z and autosomes. Other focal species result in clustering of the data into two groups, thereby making comparisons of the slope meaningless. Therefore, we calculated ρ between each species and the pheasant, and then plotted ρ for autosomal and Z genes for each expression class by divergence time. Our results show a greater rate of decline in ρ over time for the Z chromosome compared with the autosomes, consistent with Fast-Z evolution of gene expression; however, the effect was primarily observed in females (fig. 1). For genes expressed in the female spleen, ρ decreased more rapidly over time (resulting in a significantly steeper slope) for the Z chromosome compared with the autosomes (fig. 1*A*, supplementary fig. S1 and table S1, Supplementary Material online). In the male spleen, the effect was marginally nonsignificant (fig. 1*B* and table S1, Supplementary Material online). Similarly, in the gonad the slope was greater for the Z chromosome than the autosomes in females (fig. 1*C* and table S1, Supplementary Material online), but not in males (fig. 1*D* and table S1, Supplementary Material online). In the female and male spleens, ρ was significantly lower on the Z chromosome than on the autosomes in the majority of comparisons (fig. 1). In the female gonad, there was a significantly lower ρ on the Z chromosome compared with the autosomes only in comparisons between waterfowl and pheasant (fig. 1).

We also looked for signatures of Fast-Z evolution using expression divergence in all pairwise comparisons between the six species (Meisel et al. 2012). In the female spleen, we detected higher gene expression divergence for the Z chromosome than autosomes in 14 of 15 comparisons (fig. 2). In the male spleen, gene expression divergence for the Z chromosome was significantly greater than that of the autosomes for 7 of 15 pairwise comparisons (fig. 2).

In the female gonad, 14 of 15 pairwise comparisons showed higher divergence on the Z chromosomes than autosomes (fig. 3). In the male gonad, only 4 of the 15 pairwise comparisons showed higher divergence on the Z (fig. 3), and interestingly all of these comparisons with higher gene expression divergence on the Z involved divergence from duck.

Correlation of Gene Expression between Males and Females (C_{mf})

These results suggest that Fast-Z evolution of expression occurs primarily in females. Interestingly, the Fast-Z effect is weakly detectible in the male spleen, but not at all evident in the male gonad. This difference in Fast-Z evolution of expression in males may be the result of different levels of intersexual correlation in expression in somatic versus gonadal tissues. To explore the differences in intersexual correlation, and its possible effects on Fast-Z evolution, we measured the correlation in gene expression between males and females (here termed C_{mf}) across our six avian species. In order to control for phylogeny, we used Phylogenetic Generalized Least Squares (PGLS) in the R package Caper (R-Core-Team 2012); therefore, the strength of the correlation in expression across the six species (C_{mf}) was measured using r^2. In the spleen, expression levels between males and females are highly correlated across the clade both for genes on the autosomes and Z chromosome (fig. 4*A*, median $C_{mf\ (autosomes)} = 0.91$, median $C_{mf\ (Z\ chromosome)} = 0.86$; Wilcoxon rank sum, $P < 0.0001$). This suggests that most expression variation selected in females will also be expressed in males and may explain why both females and males show Fast-Z expression evolution in the spleen.

In contrast, the correlation between gene expression in males and females is much lower in the gonad (fig. 4*B*,

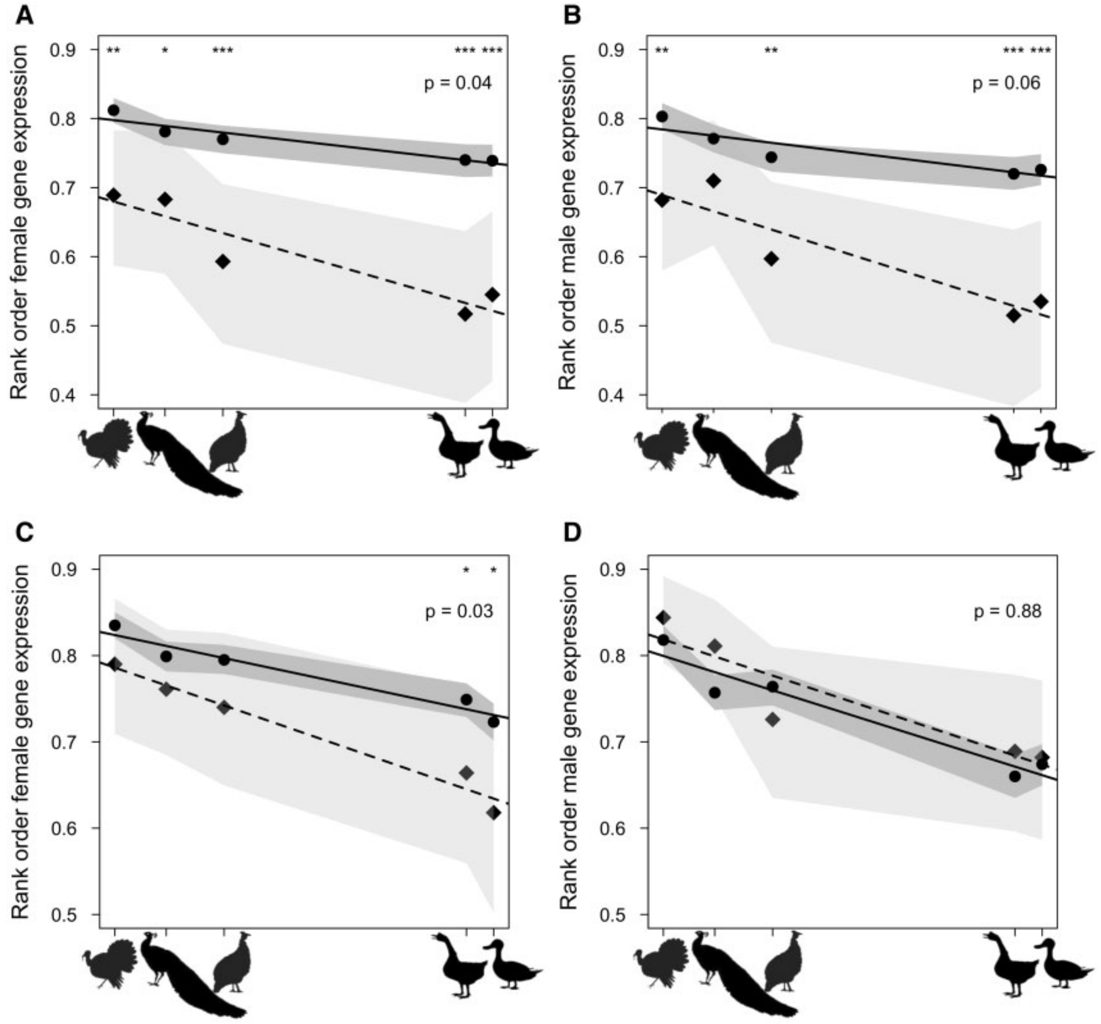


Fig. 1. Spearman's rho correlations for pairwise similarity between pheasant and each other species in the (*A*) female spleen, (*B*) male spleen, (*C*) female gonad, and (*D*) male gonad. Regression for genes on autosomes shown by solid line (and circles) and Z chromosome by dashed line (and diamonds). Shaded areas represent the 95% confidence intervals calculated through 1,000 bootstrap replicates. *P* values are for interactions between chromosome × divergence time. Significant differences between the Z chromosome and autosomes denoted by $*P < 0.05$, $**P < 0.01$, $***P < 0.001$.

$C_{mf\ (autosomes)} = 0.28$, median $C_{mf\ (Z\ chromosome)} = 0.24$; Wilcoxon rank sum, $P = 0.263$). The reduction in C_{mf} in the gonad compared with the spleen implies that most adaptive expression variation in the female gonad will not be similarly expressed in males, and may explain why Fast-Z expression evolution was only observed in females in this tissue.

Fast-Z Expression Evolution in Females Is Consistent with an Adaptive Process

The stronger signature of Fast-Z in females than males is consistent with an adaptive process driving Fast-Z evolution of gene expression due to hemizygous exposure of recessive beneficial expression variation. If Fast-Z is indeed a result of fixation of recessive beneficial alleles in females, we would expect to see greater rates of Fast-Z evolution for female-biased genes than male-biased genes. Consistent with this, we find indications of Fast-Z evolution for female-biased genes in the female gonad but not for male-biased genes in the male gonad (supplementary fig. S1, Supplementary Material online). Interestingly, both Z and autosomal female-biased loci expressed in the male gonad exhibit greater variation in divergence, with a high overall average, a pattern not observed for male-biased genes expressed in the female gonad (supplementary fig. S1, Supplementary Material online).

Additionally, if purifying selection acting on coding sequence constrains adaptive Fast-Z evolution in coding sequence, we might expect greater signatures of adaptive Fast-Z expression evolution for highly expressed genes than lowly expressed genes, as the sequence of highly expressed genes has been shown to be subject to stronger purifying selection (Resch et al. 2007). Consistent with this prediction, we find more pronounced Fast-Z expression evolution in females for genes that are highly expressed compared with those with lower expression, although we do detect signatures of Fast-Z expression evolution for both expression categories (supplementary fig. S2, Supplementary Material online). As expected, highly expressed genes in general tend to be more constrained and generally show overall lower divergence than genes that have low expression across the

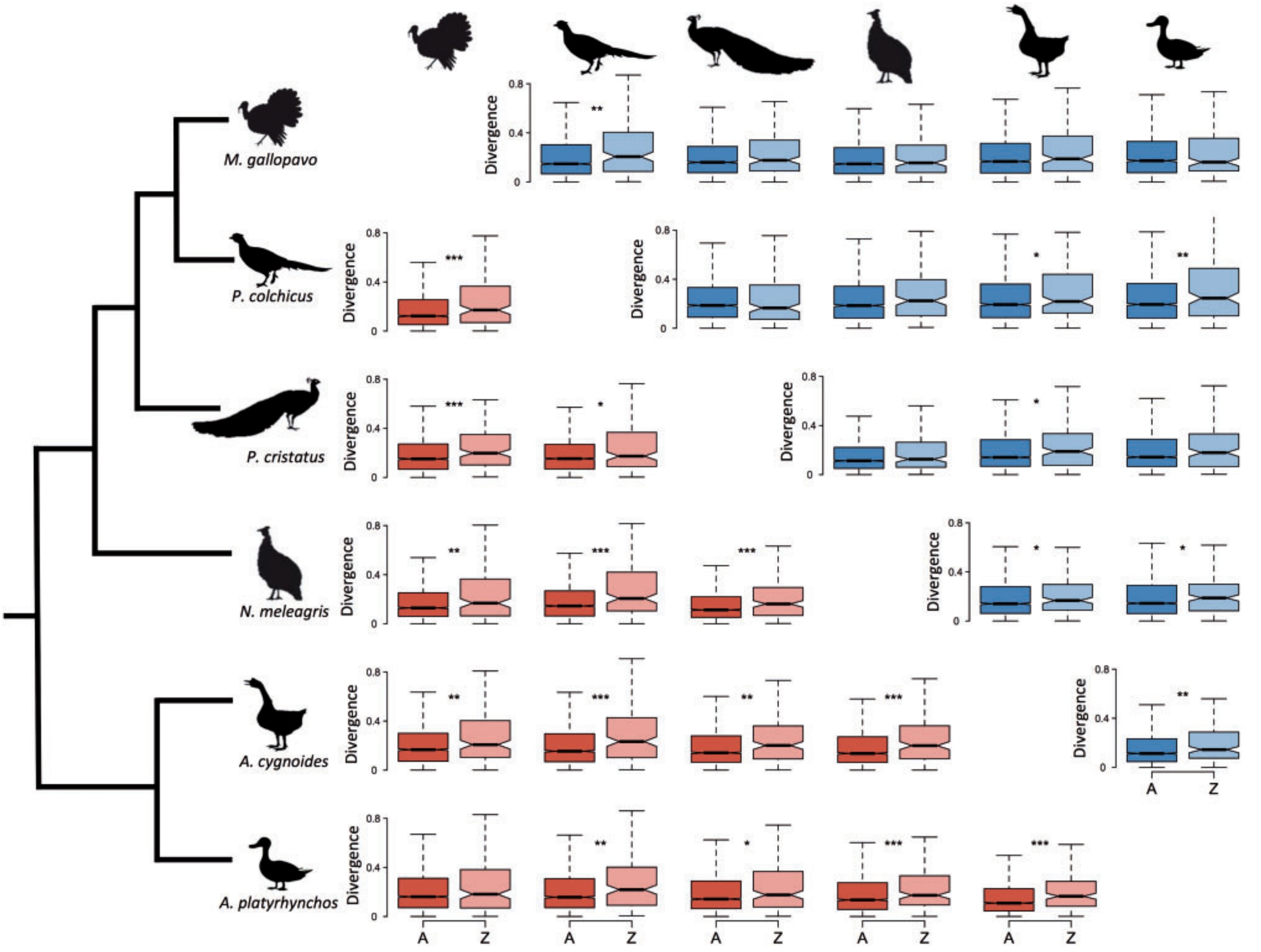


FIG. 2. Branch-specific pairwise gene expression divergence for female and male spleens. Gene expression divergence in female spleen shown below the diagonal (in red) and male spleen above the diagonal (in blue). Genes on autosomes are shaded darker and genes on Z chromosome shaded lighter. Two-sided Wilcoxon tests for significant differences between autosomal and Z chromosome divergence denoted by $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.

genome (supplementary fig. S2, Supplementary Material online).

Expression Variance Indicates Fast-Z Is due to Adaptive Evolution

In order to test more directly whether gene expression changes are due to adaptive versus nonadaptive processes, we used Δx, a measure of adaptive change in expression evolution (Moghadam et al. 2012) which incorporates both divergence and polymorphism (expression variance). We reconstructed ancestral expression levels using a maximum-likelihood (ML) estimator of Brownian Motion (Schluter et al. 1997; Paradis et al. 2004; Harrison et al. 2015). It is important to note that models of gene expression evolution are largely additive, and are not yet possible to functionally validate. Their utility in extrapolating evolutionary signals is important, but results must be interpreted cautiously. More importantly, error increases over phylogenetic space; therefore, we confine our analyses using ancestral reconstructions to the internal nodes nearest to each of our study species (nearest ancestor).

We measured gene expression divergence between each of our species and the reconstructed gene expression of the nearest ancestor (nearest internal node). In the spleen, higher gene expression divergence on the Z chromosome was in general detected for both males and females (fig. 5A). Consistent with the pairwise species comparisons in the gonad (fig. 3) we found higher expression divergence for genes on the Z than for autosomal genes in all six species comparisons in females, but not in males (fig. 6A). We calculated the proportion of genes on the Z and autosomes where divergence exceeds polymorphism ($-1 > \Delta x > 1$), a signal of positive selection (Moghadam et al. 2012). In the female spleen there was a higher proportion of genes on the Z chromosome showing putative positive selection in only one of the species, and a significantly lower proportion for one species in males (fig. 5B). However, in the female gonad in three of the six species we found a higher proportion of genes on the Z chromosome showed evidence of putative positive selection in gene expression compared with autosomes (fig. 6B). In contrast, there was no significant difference in the proportion of genes under positive selection on the Z or autosomes in any species in the male gonad (fig. 6B).

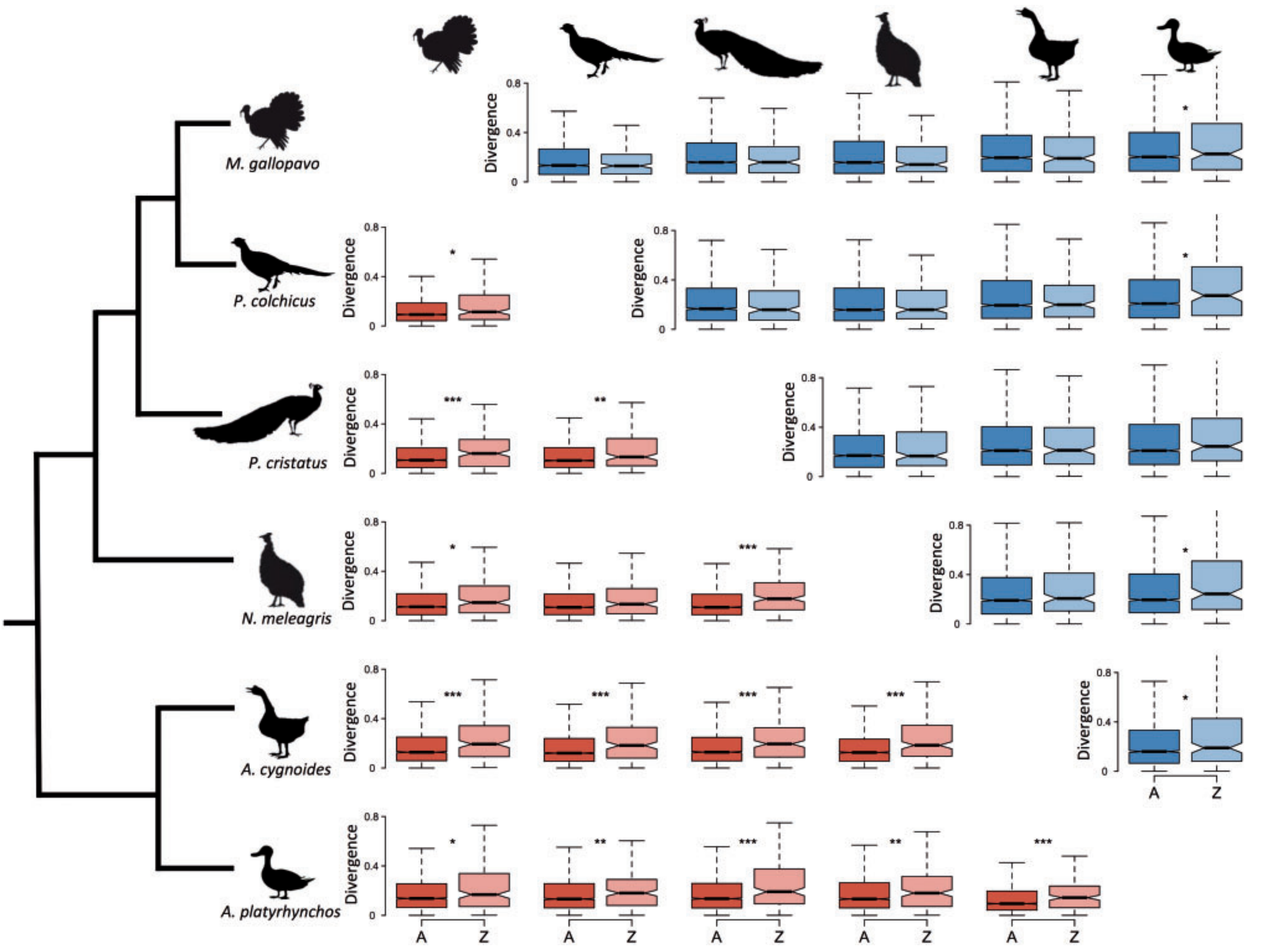


Fig. 3. Branch-specific pairwise gene expression divergence for female and male gonad. Gene expression divergence in female gonad shown below the diagonal (in red) and male gonad above the diagonal (in blue). Genes on autosomes are shaded darker and genes on Z chromosome shaded lighter. Two-sided Wilcoxon tests for significant differences between autosomal and Z chromosome divergence denoted by *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$.

Discussion

Our study finds clear signatures of Fast-Z evolution of gene expression in both the somatic and gonadal tissues, similar to a recent study on Fast-Z in gene sequence (Wright et al. 2015). However, in contrast to previous studies of protein coding data, which support a predominant role of drift in Fast-Z (Wright et al. 2015), our data indicate that Fast-Z in gene expression is primarily the result of positive selection acting in females due to hemizygous exposure of recessive beneficial variation.

Fast-Z in Gene Expression Is Largely the Result of Adaptive Evolution in Females

Our results provide several lines of evidence that support the role for positive selection in driving Fast-Z evolution of gene expression. First, the Fast-Z effect in expression is stronger in females than males, consistent with hemizygous exposure of beneficial variation. In gonadal tissue, we find strong signatures of Fast-Z in females but not males (fig. 3), and the Fast-Z effect is stronger for females than males in the spleen (fig. 2). Additionally, we find tentative support that female-biased genes show stronger Fast-Z expression evolution than male-biased genes (supplementary fig. S1, Supplementary Material online), consistent with the assumption that female-biased genes encode female phenotypes (Connallon and Clark 2011). Different methods to identify sex-biased gene expression can yield different results (Assis et al. 2012). However, our method of defining sex-bias was broadly consistent with the EdgeR method (Robinson et al. 2010), producing an overlap in sex-biased expression of 89–96% between both approaches (Wright et al. 2015). This means that our analyses of gene expression divergence for sex-biased genes are unlikely to be affected by different methods of classifying sex-biased gene expression. Finally, in females, a higher proportion of genes on the Z in several of the six species studied show evidence of positive selection for expression (figs. 5B and 6B), but we find no such difference in males. Although differences in gene function between the autosomes and Z chromosome could contribute to Fast-Z, Gene Ontology analysis for the orthologous genes across these six species suggests no difference in gene function across the autosomes and Z chromosome (Wright et al. 2015). These results taken as a whole are consistent with an adaptive explanation of Fast-Z.

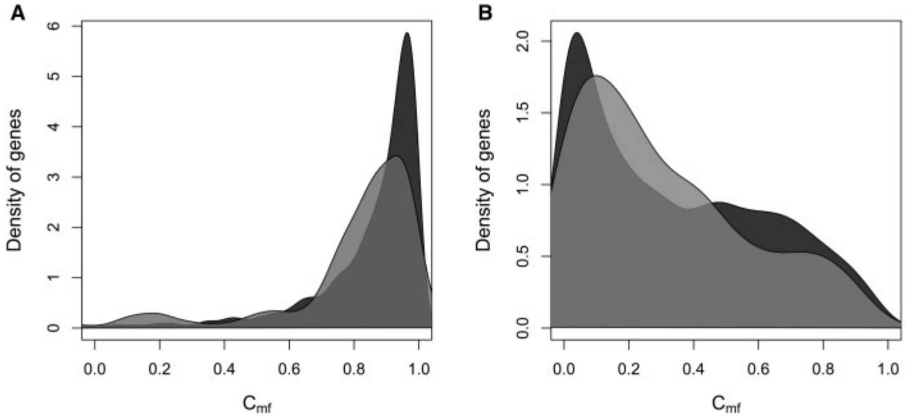


Fig. 4. Density distribution of correlations in gene expression between males and females (C_{mf}) for orthologous genes expressed in (*A*) spleen and (*B*) gonad. Correlations are r^2 values from phylogenetically controlled generalized least square models. Genes on autosomes are dark gray and on Z chromosome are light gray.

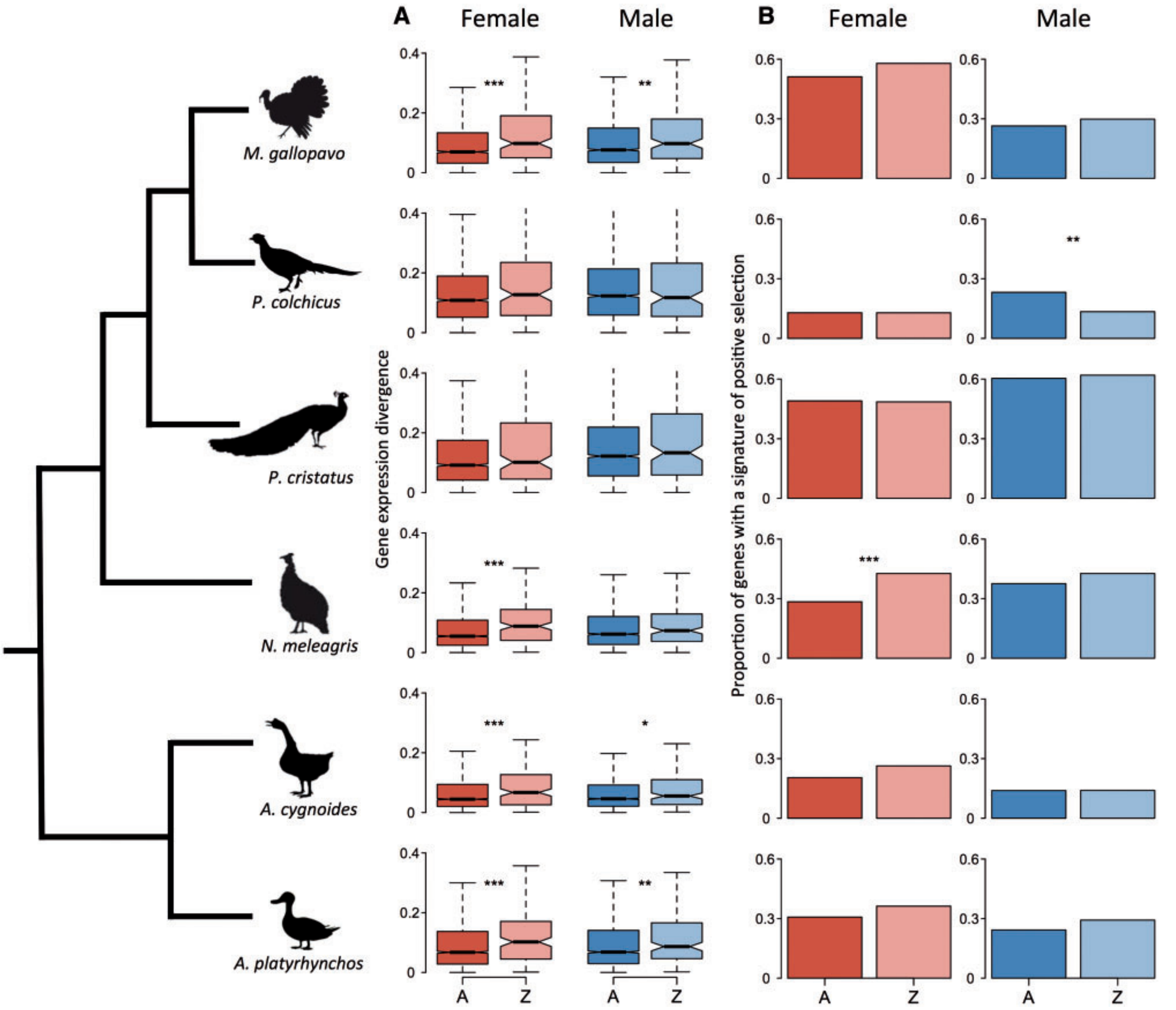


Fig. 5. (*A*) Pairwise gene expression divergence between each focal species and the estimated ancestral gene expression levels at the nearest node in female and male spleens. Two-sided Wilcoxon tests denote significant differences between autosomal and Z chromosome divergence. (*B*) Proportion of genes on the autosomes and Z chromosome with a signature of positive selection ($-1 > \Delta X > 1$) for the female and male spleens. Pearson's chi squared tests denote significant differences in the proportion of genes positively selected on Z chromosomes and autosomes. Females are on left (in red) and males on right (in blue). Autosomes are shaded dark and Z chromosome shaded light. $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.

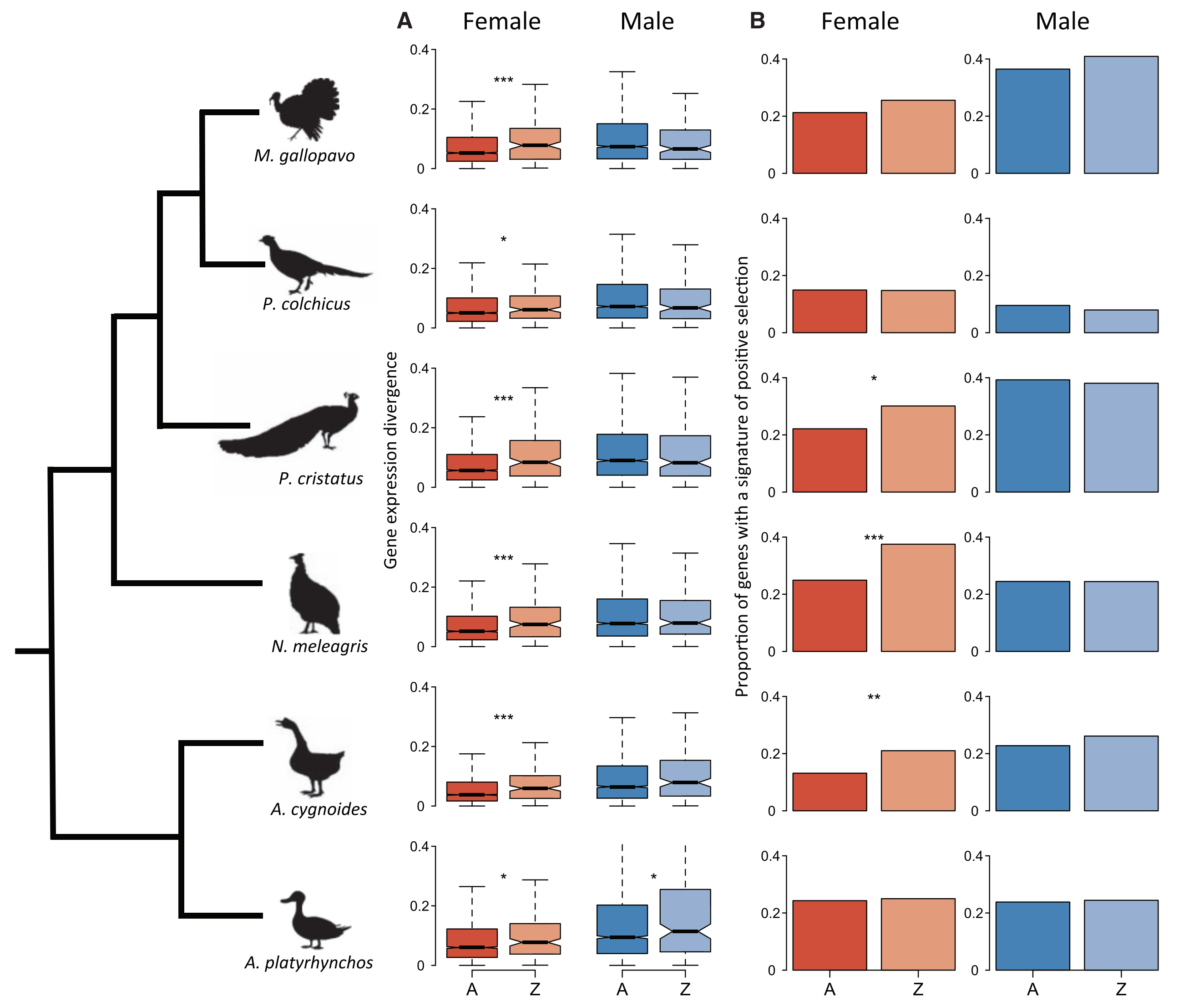


Fig. 6. (*A*) Pairwise gene expression divergence between each focal species and the estimated ancestral gene expression levels at the nearest node in the female and male gonad. Two-sided Wilcoxon tests denote significant differences between autosomal and Z chromosome divergence. (*B*) Proportion of genes on the autosomes and Z chromosome with a signature of positive selection ($-1 > \Delta x > 1$) for the female and male gonad. Pearson's chi-squared tests denote significant differences in the proportion of genes positively selected on Z chromosomes and autosomes. Females are in red and males in blue. Autosomes are shaded dark and Z chromosome shaded light. $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.

The role of selection in driving Fast-Z evolution of gene expression is perhaps surprising given that drift has been shown to be the primary cause of Fast-Z evolution in birds (Wright et al. 2015). This suggests that gene sequence and gene expression are subject to different evolutionary forces. The alternative reasons for Fast-Z gene sequence and gene expression evolution are likely to be linked to how selection acts on *cis*-regulatory regions. Mutations in *cis*-regulatory regions of genes are thought to be particularly important for evolutionary change (Wray 2007), and *cis*-regulated expression may be subject to stronger positive selection (Emerson et al. 2010) even in the face of pleiotropic constraint imposed on genes with conserved expression (Wray 2007). In contrast, the sequence for conserved orthologs may be largely shaped through purifying selection, limiting adaptive potential. Together, this suggests that the adaptive potential of Z chromosome gene expression may be greater than that of coding sequence, which may be important for studies of speciation and sexual selection, where the Z chromosome is often theoretically implicated as a major contributor (Haldane 1922; Kirkpatrick and Hall 2004).

Differences between the Spleen and Gonad in Fast-Z Evolution of Male Expression

Expression data are arguably more useful for studies of Fast-X or Fast-Z evolution because they can be used effectively to compare the sexes, as opposed to coding sequence data, which are the same in both males and females. Our analysis shows that Fast-Z expression evolution is consistent in the female gonad and soma, but is only weakly detectible in the male spleen and is absent from the male gonad (figs. 1–3). The difference in male Fast-Z expression evolution between the spleen and the gonad may be a consequence of the difference in the strength of the genetic correlation between the sexes in these different tissues.

In the spleen, expression in males and females is highly correlated across the phylogeny (fig. 4*A*); therefore, the fixation of expression variation on the Z chromosome in females will often also result in the same expression pattern in males, producing a weaker, but still detectible, signature of Fast-Z expression evolution in the male spleen. In contrast to the spleen, the genetic correlation (C_{mf}) in expression between males and females is much lower in gonadal tissue (fig. 4*B*). This suggests that the majority of expression variation in the gonad is sex-specific in its effects, and therefore fixation of expression variants that are beneficial to females on the Z chromosome would not necessarily result in the same pattern when expressed in males. We also note that differences in the intersexual genetic correlation for genes on the Z chromosome and autosome are unlikely to contribute to our patterns of Fast-Z expression evolution because there were only small but significant differences in C_{mf} between the autosomes and Z chromosome in the spleen and there was no significant difference in the gonad.

Another important difference between somatic and gonadal tissue is the extent of dosage compensation. In birds, there is generally a lack of dosage compensation in the gonad, whereas the spleen tends to exhibit a degree of incomplete dosage compensation (Ellegren et al. 2007; Itoh et al. 2007). Differences in the extent of dosage compensation are thought to affect Fast-Z sequence evolution due to beneficial mutations (Mank, Vicoso, et al. 2010). When dosage compensation is more complete, Fast-Z sequence evolution due to positive selection is thought to be more pronounced, potentially because the selection coefficients in the heterogametic sex are expected to be smaller (Charlesworth et al. 1987). However, contrary to this, our data show similar patterns of Fast-Z expression evolution in the female gonad and spleen. As we do not see variation in the magnitude of dosage compensation across the six species studied, selection for dosage compensation is unlikely to drive the Fast-Z effect that we detect.

Fast-X versus Fast-Z

Faster rates of gene expression evolution on sex chromosomes have been detected in mammals and *Drosophila*, both male heterogametic systems. In mammals, the evidence suggests that Fast-X evolution of gene expression occurred as an adaptive burst on the newly formed therian X (Brawand et al. 2011). Similarly, we also find signatures of Fast-Z in expression over short evolutionary timescales (i.e., between closely related species, figs. 2 and 3), and at the tips of the phylogenetic tree (figs. 5*A* and 6*A*). However, in contrast to the mammalian study, we also find Fast-Z across more distantly related species (figs. 2 and 3), and the level of Fast-Z is correlated with phylogenetic distance (fig. 1), suggesting that the effect is cumulative over time.

Studies on Fast-X in *Drosophila* have shown that Fast-X is more strongly detected, but not limited to, male-biased genes expressed in male reproductive tissue (Meisel et al. 2012), although another study showed that Fast-X was restricted to *Drosophila* embryonic stages (Kayserili et al. 2012). Both studies are consistent with Fast-X driven by the adaptive fixation of mutations that affect gene expression in *cis*. Our data on Fast-Z provide further support that mutations affecting gene expression of genes on sex chromosomes are also primarily regulated in *cis*, and that the fitness consequences of these mutations are in general recessive (Meisel et al. 2012). Furthermore, *Drosophila* exhibits complete X chromosome dosage compensation and Z chromosome dosage compensation in birds is incomplete (reviewed in Mank 2013). The similarity between expression Fast-X in *Drosophila* and Fast-Z in birds suggests that faster rates of gene expression evolution are not restricted to a particular mode of dosage compensation (Meisel et al. 2012).

Models of Gene Expression Divergence

We note that measuring Fast-Z using gene expression rather than gene sequence may present a few caveats. First, current models of gene expression evolution assume additivity (Brawand et al. 2011; Ometto et al. 2011; Moghadam et al. 2012; Rohlfs et al. 2013), which has yet to be validated (Khaitovich et al. 2006). Second, in species with incomplete dosage compensation such as birds (Mank 2013), genes on sex chromosomes in the heterogametic sex will often have lower expression than genes on autosomes, which may affect measures of Fast-Z. However, our measure of gene expression divergence takes into account expression level and so this should not affect our ability to detect Fast-Z. Furthermore, we detect Fast-Z for both highly and lowly expressed genes, and our results are robust to different measures of Fast-Z, such as Spearman's rho correlation coefficient and gene expression divergence calculations.

Final Remarks

We detect Fast-Z evolution in gene expression across six avian species spanning 90 My of evolutionary history, and our results indicate that, in contrast to studies of coding sequence, Fast-Z in expression is primarily due to adaptive evolution of female-benefit variation. Together, this suggests that the adaptive potential of Z chromosome gene expression may be greater than that of coding sequence, which may be important for studies of speciation and sexual selection, where the Z chromosome has been theoretically shown to play a major role (Haldane 1922; Kirkpatrick and Hall 2004).

Materials and Methods

Transcriptome Assembly

Spleen and gonad samples were collected from captive-reared males and females at the start of their first breeding season for *Anas platyrhynchos* (mallard duck), *Meleagris gallopavo* (wild turkey), *Phasianus colchicus* (common pheasant), *Numida meleagris* (helmeted guineafowl), *Pavo cristatus* (Indian peafowl) and *Anser cygnoides* (swan goose), with permission from institutional ethical review committees and in accordance with national guidelines.

The left gonad and spleen were dissected separately from five males and five females for *A. platyrhynchos*, *N. meleagris*, *P. cristatus* and *A. cygnoides*, and from six males and five females for *P. colchicus*. In *M. gallopavo*, four male and two female spleens were collected and five male and female gonads were collected. Samples were homogenized, stored initially in RNAlater, and RNA was then prepared with the Animal Tissue RNA Kit (Qiagen). mRNA was subtracted and individual samples barcoded by The Wellcome Trust Centre for Human Genetics, University of Oxford using Illumina's Multiplexing Sample Preparation Oligonucleotide Kit with an insert size of 280 bp. RNA was sequenced on an Illumina HiSeq 2000 resulting in on average 26 million 100-bp paired-end reads per sample.

Quality control, de novo assembly, and ortholog detection have been described previously (Harrison et al. 2015; Wright et al. 2015). Reads were mapped to de novo assemblies to obtain expression levels. Comparisons of normalized expression counts were used to identify sex-biased gene expression using standard measures and corrected for multiple testing (Pointer et al. 2013; Perry et al. 2014).

Genes used in all subsequent analyses were restricted to reciprocal 1–1 orthologs across all six study species that were expressed in either sex. We filtered out any sex-limited gene with expression less than 2 rpkm in the sex in which it was expressed, then removed any genes that were not expressed in all six of the species, resulting in 2,428 autosomal genes and 171 Z-linked genes for the spleen, and 2,729 autosomal and 184 Z-linked genes for the gonad. Analyses of gene expression similarity and divergence were done for males and females separately in R v.2.15.1 (R-Core-Team 2012).

Divergence and Phylogeny Estimation

In order to estimate divergence time as well as phylogenetic distance, nucleotide sequences for reciprocal orthologous genes were aligned with PRANK v.130820 (Löytynoja and Goldman 2008) using ML-derived guide trees, with the zebra finch as an outgroup. Reciprocal orthologs were used to construct an ML phylogeny for our six species with a GBLOCKS 0.91b (Castresana 2000) filtered alignment using RaxML (Stamatakis 2014) version 7.4.2. The gene set was filtered with Repeatmasker (http://www.repeatmasker.org) to remove retrotransposons and tandem repeats. Genes were also checked for in-frame internal stop codons and SWAMP version 1.0 (Harrison et al. 2014) was used with a threshold of four in a window size of five bases to check for regions with poor alignment and to set a minimum sequence length of 75 bp. PAML version 4.7a (Yang 2007) was used to estimate divergence for orthologous genes, and orthologous genes with $d_S > 2$ were removed from further analyses as this represents the point of mutational saturation in avian sequence data (Axelsson et al. 2008). The resulting molecular divergence was measured as root-to-tip branch length between pheasant and each species.

Measures of Fast-Z Evolution of Gene Expression

Spearman's Rho

Spearman's rho correlation coefficient (ρ) can be used to estimate the decay in similarity between species over time, and the comparison between the slope of ρ across phylogenetic distance for the Z and autosomes is a measure of Fast-Z evolution of gene expression. Spearman's rho correlation between pheasant and all other species was calculated for all genes on the autosomes and Z chromosome (Kayserili et al. 2012; Harrison et al. 2015). We used linear models to test for a significant difference between the slope of the decay in similarity across molecular divergence time for autosomes and Z chromosome (supplementary table S1, Supplementary Material online). For each pairwise comparison, we tested whether the Z chromosome ρ was significantly different from the autosomal ρ using 1,000 bootstrapped replicates consisting of the number of Z-linked genes sampled from the pool of autosomal genes. The 95% confidence intervals of the autosomal distribution were used to denote a significant difference between the Z chromosome and the autosomes.

Gene Expression Divergence

For each pairwise species comparison, expression divergence was calculated as the difference in gene expression between the two species divided by the average gene expression (Meisel et al. 2012) for each locus. Gene expression divergence was calculated separately for male and female expression in the spleen and gonad. Two-sided Wilcoxon tests were used to test for differences in gene expression divergence on autosomes and Z chromosomes.

Correlation in Gene Expression between Males and Females

The correlation in gene expression between males and females (C_{mf}) was calculated separately for each gene for expression in the spleen and gonad. PGLS models were used in the Caper package (Orme et al. 2012) (R v2.15.1) using the ML phylogeny for our six species to correct for phylogeny. The r^2 value was used as the estimate of the strength of the correlation in gene expression between males and females for each gene. Sex limited genes were removed from these analyses as the models cannot account for low variance in expression across the phylogeny.

Ancestral State Gene Expression Divergence and Directional Selection

Ancestral state reconstruction of expression was conducted with the APE package (Paradis et al. 2004) using the Brownian motion-based ML estimator (Schluter et al. 1997) using the ML phylogeny of the six species described above. Gene expression divergence was calculated between each species and their most recent ancestor (i.e., their nearest internal node in the phylogenetic tree).

Models exist to test for positive selection in gene expression (Brawand et al. 2011; Roux et al. 2014). Δx (Moghadam et al. 2012) is particularly useful in this case because it corrects for expression level, which is important in comparisons between diploid and haploid chromosomes, and in systems lacking complete sex chromosome dosage compensation. We calculated Δx (Moghadam et al. 2012) between each species and their most recent ancestor (i.e., their nearest internal node in the phylogenetic tree). Δx incorporates expression variance as an indicator of polymorphism, and values for $\Delta x > 1$ or < -1 indicate that divergence from the point estimate of the ancestral state is greater than standing genetic variation in gene expression within the species, a typical indicator of positive selection.

Sex-Biased Gene Expression

Sex-biased gene expression rapidly changes across the phylogeny (Harrison et al. 2015) and few genes remain sex-biased in all six species (Harrison et al. 2015). Genes whose ancestral reconstruction was sex-biased at the ancestral node to all six species were therefore classified as sex-biased. Male- and female-biased genes were identified using log 2-fold change gene expression between males and females. This resulted in 24 female-biased genes and 54 male-biased genes on the Z chromosome and 589 female-biased genes and 554 male-biased genes on the autosomes.

Expression Level

Genes were broadly divided into highly and lowly expressed. Average gene expression across the six species was calculated for each gene and then expression was averaged across males and females. Genes with expression above the median (4.55 rpkm) were classified as highly expressed and below were classified as lowly expressed. Significant differences in gene expression divergence on the Z chromosome and autosomes between the ancestral state and each species were analyzed as before.

Supplementary Material

Supplementary figures S1 and S2 and table S1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Stephen Montgomery, Natasha Bloch and two anonymous reviewers for helpful comments and suggestions, as well as Marie Pointer for help with sample collection. We acknowledge the use of the University College London Legion High Performance Computing Facility (Legion@UCL), Unity Shared Memory Facility, and associated support services in the completion of this work. Sequencing was performed by the Oxford University Wellcome Trust Centre for Human Genetics (funded by Wellcome Trust Grant Reference 090532/Z09/Z and Medical Research Council Hub Grant G0900747 91070). This work was supported by the European Research Council under the Framework 7 Agreement (Grant Agreement 260233 to J.E.M.).

References

Andersson M. 1994. Sexual selection. Princeton (NJ): Princeton University Press.

Assis R, Zhou Q, Bachtrog D. 2012. Sex-biased transcriptome evolution in *Drosophila*. *Genome Biol Evol.* 4:1189–1200.

Axelsson E, Hultin-Rosenberg L, Brandstrom M, Zwahlen M, Clayton DF, Ellegren H. 2008. Natural selection in avian protein-coding genes expressed in brain. *Mol Ecol.* 17:3008–3017.

Brawand D, Soumillon M, Necsulea A, Julien P, Csardi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, et al. 2011. The evolution of gene expression levels in mammalian organs. *Nature* 478:343–348.

Caballero A. 1995. On the effective size of populations with separate sexes, with particular reference to sex-linked genes. *Genetics* 139:1007–1011.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol.* 17:540–552.

Charlesworth B, Coyne JA, Barton NH. 1987. The relative rates of evolution of sex chromosomes and autosomes. *Am Nat.* 130:113–146.

Connallon T, Clark AG. 2011. Association between sex-biased gene expression and mutations with sex-specific phenotypic consequences in Drosophila. *Genome Biol Evol.* 3:151–155.

Corl A, Ellegren H. 2012. The genomic signature of sexual selection in the genetic diversity of the sex chromosomes and autosomes. *Evolution* 66:2138–2149.

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

Emerson JJ, Hsieh LC, Sung HM, Wang TY, Huang CJ, Lu HH, Lu MY, Wu SH, Li WH. 2010. Natural selection on cis and trans regulation in yeasts. *Genome Res.* 20:826–836.

Haldane JBS. 1922. Sex-ratio and unisexual sterility in hybrid animals. *J Genet.* 12:101–109.

Harrison PW, Jordan GE, Montgomery SH. 2014. SWAMP: Sliding Window Alignment Masker for PAML. *Evol Bioinform Online.* 10:197–204.

Harrison PW, Wright AE, Zimmer F, Dean R, Montgomery S, Pointer MA, Mank JE. 2015. Sexual selection drives evolution and rapid turnover of male gene expression. *Proc Natl Acad Sci U S A.* 112:4393–4398.

Itoh Y, Melamed E, Yang X, Kampf K, Wang S, Yehya N, van Nas A, Replogle K, Band MR, Clayton DF, et al. 2007. Dosage compensation is less effective in birds than in mammals. *J Biol.* 6:2.

Kayserili MA, Gerrard DT, Tomancak P, Kalinka AT. 2012. An excess of gene expression divergence on the X chromosome in *Drosophila* embryos: implications for the faster-X hypothesis. *PLoS Genet.* 8:e1003200.

Khaitovich P, Enard W, Lachmann M, Paabo S. 2006. Evolution of primate gene expression. *Nat Rev Genet.* 7:693–702.

Kirkpatrick M, Hall DW. 2004. Male-biased mutation, sex linkage, and the rate of adaptive evolution. *Evolution* 58:437–440.

Laporte V, Charlesworth B. 2002. Effective population size and population subdivisions in demographically structured populations. *Genetics* 162:501–519.

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

Mank JE. 2013. Sex chromosome dosage compensation: definitely not for everyone. *Trends Genet.* 29:677–683.

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

Mank JE, Nam K, Ellegren H. 2010. Faster-Z evolution is predominantly due to genetic drift. *Mol Biol Evol.* 27:661–670.

Mank JE, Vicoso B, Berlin S, Charlesworth B. 2010. Effective population size and the Faster-X effect: empirical evidence and its interpretation. *Evolution* 64:663–674.

Meiklejohn CD, Coolon JD, Hartl DL, Wittkopp PJ. 2014. The roles of cis and trans-regulation in the evolution of regulatory incompatibilities and sexually dimorphic gene expression. *Genome Res.* 24:84–95.

Meisel RP, Connallon T. 2013. The faster-X effect: integrating theory and data. *Trends Genet.* 29:537–544.

Meisel RP, Malone JH, Clark AG. 2012. Faster-X evolution of gene expression in *Drosophila*. *PLoS Genet.* 8:e1003013.

Moghadam HK, Pointer MA, Wright AE, Berlin S, Mank JE. 2012. W chromosome expression responds to female-specific selection. *Proc Natl Acad Sci U S A.* 109:8207–8211.

Ometto L, Shoemaker DW, Ross KG, Keller L. 2011. Evolution of gene expression in fire ants: the effects of developmental stage, caste, and species. *Mol Biol Evol.* 28:1381–1392.

Orme D, Freckleton R, Thomas G, Petzoldt T, Fritz S, Isaac N, Pearse W. 2012. caper: comparative analyses of phylogenetics and evolution in R.

Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 21:289–290.

Perry JC, Harrison PW, Mank JE. 2014. The ontogeny and evolution of sex-biased gene expression in *Drosophila melanogaster*. *Mol Biol Evol.* 31:1206–1219.

Pointer MA, Harrison PW, Wright AE, Mank JE. 2013. Masculinization of gene expression is associated with exaggeration of male sexual dimorphism. *PLoS Genet.* 9:e1003697.

R-Core-Team. 2012. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.

Resch AM, Carmel L, Marino-Ramirez L, Ogurtsov AY, Shabalina SA, Rogozin IB, Koonin EV. 2007. Widespread positive selection in synonymous sites of mammalian genes. *Mol Biol Evol.* 24:1821–1831.

Robinson MD, McCarthy DJ, Smyth GK. 2010. EdgeR a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26:139–140.

Rohlfs RV, Harrigan P, Nielsen R. 2013. Modeling gene expression evolution with an extended Ornstein-Uhlenbeck process accounting for within-species variation. *Mol Biol Evol.* 31:201–211.

Roux J, Privman E, Moretti S, Daub JT, Robinson-Rechavi M, Keller L. 2014. Patterns of positive selection in seven ant genomes. *Mol Biol Evol.* 31:1661–1685.

Sackton TB, Corbett-Detig RB, Nagaraju J, Vaishna L, Arunkumar KP, Hartl DL. 2014. Positive selection drives faster-Z evolution in silkmoths. *Evolution* 68:2331–2342.

Schluter D, Price T, Mooers AØ, Ludwig D. 1997. Likelihood of ancestor states in adaptive radiation. *Evolution* 51:1699–1711.

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

Stern DL, Orgogozo V. 2008. The loci of evolution: how predictable is genetic evolution? *Evolution* 62:2155–2177.

Vicoso B, Charlesworth B. 2006. Evolution on the X chromosome: unusual patterns and processes. *Nat Rev Genet.* 7:645–953.

Vicoso B, Charlesworth B. 2009. Effective population size and the Faster-X effect: an extended model. *Evolution* 63:2413–2426.

Vicoso B, Emerson JJ, Zektser Y, Mahajan S, Bachtrog D. 2013. Comparative sex chromosome genomics in snakes: differentiation, evolutionary strata, and lack of global dosage compensation. *PLoS Biol.* 11:e1001643.

Wade MJ. 1979. Sexual selection and variance in reproductive success. *Am Nat.* 114:742–747.

Wang HY, Chien HC, Osada N, Hashimoto K, Sugano S, Gojobori T, Chou CK, Tsai SF, Wu CI, Shen CK. 2007. Rate of evolution in brain-expressed genes in humans and other primates. *PLoS Biol.* 5:335–342.

Wang Z-J, Zhang J, Yang W, An N, Zhang P, Zhang G-J, Zhou Q. 2014. Temporal genomic evolution of bird sex chromosomes. *BMC Evol Biol.* 14:250.

Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet.* 8:206–216.

Wright AE, Harrison PW, Zimmer F, Montgomery S, Pointer MA, Mank JE. 2015. Variation in promiscuity and sexual selection drives avian rate of Faster-Z evolution. *Mol Ecol.* 24:1218–1235.

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

The following article was first published in *Genetics*

Wright AE, **Zimmer F**, Harrison PW, Mank JE (2015) Conservation of regional variation in sex-specific sex chromosome regulation. *Genetics* 201: 587-598
doi: 10.1534/genetics.115.179234

The following article was first published in

*Molecular Ecolog*y

Wright AE, Harrison PW, **Zimmer F**, Montgomery SH, Pointer MA, Mank JE (2015) Variation in promiscuity and sexual selection drives avian rate of Faster-Z evolution. *Molecular Ecolog*y 24:1218-1235
doi: 10.1111/mec.13113

MOLECULAR ECOLOGY

Molecular Ecology (2015) **24**, 1218–1235 doi: 10.1111/mec.13113

Variation in promiscuity and sexual selection drives avian rate of Faster-Z evolution

ALISON E. WRIGHT,*† PETER W. HARRISON,† FABIAN ZIMMER,† STEPHEN H. MONTGOMERY,† MARIE A. POINTER* and JUDITH E. MANK†

**Department of Zoology, Edward Grey Institute, University of Oxford, Oxford OX1 3PS, UK, †Department of Genetics, Evolution and Environment, University College London, London WC1E 6BT, UK*

Abstract

Higher rates of coding sequence evolution have been observed on the Z chromosome relative to the autosomes across a wide range of species. However, despite a considerable body of theory, we lack empirical evidence explaining variation in the strength of the Faster-Z Effect. To assess the magnitude and drivers of Faster-Z Evolution, we assembled six *de novo* transcriptomes, spanning 90 million years of avian evolution. Our analysis combines expression, sequence and polymorphism data with measures of sperm competition and promiscuity. In doing so, we present the first empirical evidence demonstrating the positive relationship between Faster-Z Effect and measures of promiscuity, and therefore variance in male mating success. Our results from multiple lines of evidence indicate that selection is less effective on the Z chromosome, particularly in promiscuous species, and that Faster-Z Evolution in birds is due primarily to genetic drift. Our results reveal the power of mating system and sexual selection in shaping broad patterns in genome evolution.

Keywords: effective population size, Faster-Z evolution, genetic drift, sexual selection

Received 25 August 2014; revision received 10 February 2015; accepted 11 February 2015

Introduction

Sex chromosomes are subject to unique evolutionary forces as a result of their unusual pattern of inheritance (Charlesworth *et al.* 1987; Vicoso & Charlesworth 2009; Connallon *et al.* 2012). The magnitude of selection, genetic drift and recombination are all predicted to differ between the sex chromosomes and autosomes (Rice 1984; Kirkpatrick & Hall 2004a; Mank *et al.* 2010a; Meisel & Connallon 2013) and studies contrasting the evolution of sex-linked to autosomal genes can shed light on the fundamental evolutionary forces acting across the genome as a whole.

Faster rates of coding sequence evolution have been observed on the Z and X chromosomes relative to the autosomes across a wide range of species (recently reviewed by Meisel & Connallon 2013), and Faster-X and Faster-Z Effects appear to be a common feature of sex chromosome evolution. However, despite elevated rates of evolution for both X-linked and Z-linked genes, the underlying causes of Faster-X and Faster-Z Evolution are predicted to differ (Vicoso & Charlesworth 2009; Meisel & Connallon 2013).

The effective population size of X and Z chromosomes (N_{EX} and N_{EZ}) is ¾ that of the autosomes (N_{EA}) when there is no difference in the variance of male and female reproductive success, such as in strictly monogamous breeding systems (Charlesworth *et al.* 1987). However, many forms of sexual selection cause elevated variance in male reproductive success (Andersson 1994), which reduces N_{EZ}/N_{EA}, and in extreme cases where a single male monopolizes the reproductive output of many females, N_{EZ} approaches ½ N_{EA} (Vicoso & Charlesworth 2009; Wright & Mank 2013) (Fig. 1). Correspondingly, genetic drift and fixation of weakly deleterious mutations is greater on the Z chromosome (Charlesworth 2009), and we predict a Faster-Z Effect largely due to neutral, nonadaptive processes. Empirical evidence in birds and snakes is consistent with this nonadaptive and neutral explanation of Faster-Z (Mank *et al.* 2010b; Corl & Ellegren 2012; Vicoso *et al.* 2013a); however, silk moths may present a recent exception

Correspondence: Alison E. Wright, E-mail: alison.e.wright@ucl.ac.uk

(Sackton *et al.* 2014). It is worth noting that a major factor determining the relative contribution of nonadaptive and adaptive drivers of Faster-Z is overall effective population size (Meisel & Connallon 2013). Overall N_E mediates the distribution of fitness effects, and specifically, we expect the efficacy of selection and adaptive component of Faster-Z to be weaker in populations with smaller N_E (Kimura & Ohta 1971).

The opposite relationship between male mating success and relative N_{EX} is predicted in male heterogametic systems (Laporte & Charlesworth 2002; Vicoso & Charlesworth 2009; Wright & Mank 2013). Increasing variance in male reproductive success results in $N_{EX}/N_{EA} > ¾$, and N_{EX}/N_{EA} may approach 1 in extreme cases (Fig. 1). Correspondingly, the higher ratio of N_{EX}/N_{EA} is expected to decrease the effect of genetic drift in Faster-X Evolution. Elevated rates of evolution on X chromosomes are therefore more often thought to be the product of increased efficacy of selection acting on recessive X-linked alleles in the heterogametic sex, thereby increasing the rate of fixation of beneficial alleles relative to the autosomes. Consistent with adaptive Faster-X Evolution, signatures of positive selection have been uncovered on the X chromosome of mammals and *Drosophila* (Thornton & Long 2005; Baines *et al.* 2008; Hvilsom *et al.* 2012; Langley *et al.* 2012).

A key prediction is that the magnitude of Faster-Z Evolution can be explained by variation in the effective population size of the sex chromosomes relative to the autosomes driven by sexual selection (Vicoso & Charlesworth 2009). Here, we explicitly test this prediction in the Galloanserae, a clade of birds spanning 90 million years (Fig. 2), for which there is extensive variation in mating system (Moller 1988, 1991; Birkhead & Petrie 1995). Using *de novo* transcriptomes for six Galloanserae species, we measured sequence divergence, polymorphism and expression and combined these molecular data with phenotypic measures of mating system to explore the nature of Faster-Z Evolution. Our results build on previous findings to reveal the dominant role nonadaptive processes play in Faster-Z. Furthermore, we uncover a positive association between Faster-Z and measures of sperm competition, a widely used indicator of the strength of postcopulatory sexual selection (Birkhead & Moller 1998). Our results suggest that variation in male mating success drives Z-linked divergence, and present the first empirical evidence in support of the considerable body of theory (Charlesworth *et al.* 1987; Vicoso & Charlesworth 2009) outlining the relationship between sexual selection and sex chromosome evolution.

Materials and methods

De novo transcriptome assembly

RNA-Seq data were obtained from captive populations of the following Galloanserae species at the start of their first breeding season; *Anas platyrhynchos* (mallard

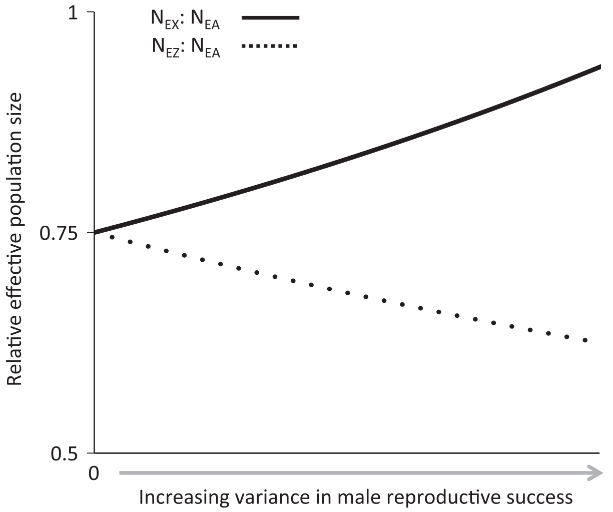


Fig. 1 Relationship between effective population size (N_E) and variance in male reproductive success. Schematic outlining the predicted relationship between variance in male reproductive success and relative N_{EZ} and N_{EX}. When variance in reproductive success is the same in males and females, under monogamy, both N_{EZ} and N_{EX} = ¾ N_{EA}. As variance in male mating success increases, N_{EZ} < ¾ N_{EA} and N_{EX} > ¾ N_{EA}.

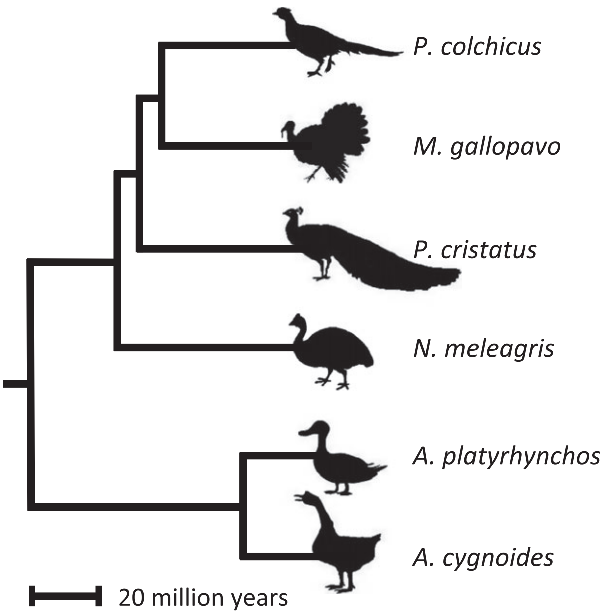


Fig. 2 Phylogenetic relationship of the Galloanserae species in this study.

duck), *Meleagris gallopavo* (wild turkey), *Phasianus colchicus* (common pheasant), *Numida meleagris* (helmeted guinea fowl), *Pavo cristatus* (Indian peafowl) *and Anser cygnoides* (swan goose) (Fig. 2). Samples were collected with permission from institutional ethical review committees and in accordance with national guidelines. The left gonad and spleen were dissected separately from five males and five females of each species. The exceptions were *P. colchicus*, where six male gonad and spleen samples were collected, and *M. gallopavo*, where four male and two female spleens were collected. Samples were homogenzied and stored in RNA later until preparation. We used the Animal Tissue RNA Kit (Qiagen) to extract RNA, and the samples were prepared and barcoded at The Wellcome Trust Centre for Human Genetics, University of Oxford using Illumina's Multiplexing Sample Preparation Oligonucleotide Kit with an insert size of 280 bp. RNA was sequenced on an Illumina HiSeq 2000 resulting in on average 26 million 100 bp paired-end reads per sample (Tables S1 and S2, Supporting Information).

The data were quality assessed using FastQC v0.10.1 (www.bioinformatics.babraham.ac.uk/projects/fastqc) and filtered using Trimmomatic v0.22 (Lohse *et al.* 2012). Specifically, we removed reads containing adaptor sequences and trimmed reads if the sliding window average Phred score over four bases was <15 or if the leading/trailing bases had a Phred score <4. Reads were removed post filtering if either read pair was <25 bases in length. We constructed *de novo* transcriptome assemblies for each species using TRINITY with default parameters (Grabherr *et al.* 2011). We separately mapped back all of the reads from each sample to the Trinity contigs using RSEM v1.1.21 with default parameters (Li & Dewey 2011) to obtain expression levels. We applied a minimum expression filter of 2 reads per kilobase per million mapped reads (RPKM) requiring that each contig has expression above unlogged 2 RPKM in at least half of any of the tissues from either sex. For each Trinity contig cluster, the isoform with the highest expression level was selected for further analysis. We removed rRNA transcripts using *G. gallus* known sequences. This generated 37453 contigs for *A. platyrhynchos*, 50817 for *M. gallopavo*, 56090 for *P. colchicus*, 45535 for *N. meleagris*, 56604 for *P. cristatus* and 44144 for *A. cygnoides*.

Identification of Galloanserae orthogroups

G. gallus (Galgal4/GCA_000002315.2) cDNA sequences were obtained from ENSEMBL v73 (Flicek *et al.* 2013), and the longest transcript for each gene was identified. We determined orthology using reciprocal BLASTN v2.2.27+ (Altschul *et al.* 1990) with an *E*-value cut-off of 1×10^{-10} and minimum percentage identity of 30%. Reciprocal 1-1 orthologs across all seven species (orthogroups) were identified using the highest BLAST score.

Avian chromosome structure is unusually stable, potentially due to a lack of active transposons (Toups *et al.* 2011), and major genomic rearrangements are infrequent (Stiglec *et al.* 2007). Synteny of the Z chromosome has previously been shown to be highly conserved across both extant birds (Vicoso *et al.* 2013b), as well as within the Galloanserae (Skinner *et al.* 2009). Chromosomal location was therefore assigned from *G. gallus* reciprocal orthologs.

Estimating sequence divergence across orthogroups

To extract Galloanserae protein-coding sequences, *G. gallus* (Galgal4/GCA_000002315.2) protein sequences were obtained from ENSEMBL v73 (Flicek *et al.* 2013). For each orthogroup, each contig was translated into all potential reading frames and BLASTED against the orthologous *G. gallus* protein sequence using BLASTX. BLASTX outputs were used to determine coding frame, and protein-coding sequences for each species were extracted. Protein-coding sequences were defined as sequences starting with the amino acid M and terminating with a stop codon or end of the contig. Orthogroups with no BLASTX hits or a valid protein-coding sequence were excluded.

Orthogroups were aligned with PRANK v121218 using the orthologous *Taeniopygia guttata* cDNA (taeGut3.2.4.75) as an outgroup and specifying the following guidetree (((*A. cygnoides*, *A. platyrhynchos*), (*N. meleagris*, (*P. cristatus*, (*M. gallopavo*, *P. colchicus*)))), *T. guttata*). Retrotransposons were removed with REPEATMASKER (v open-4.0.3), and sequences with internal stop codons were also removed. SWAMP v0.9 (Harrison *et al.* 2014) with a cut-off of 4 and window size of 15, and a minimum length of 75 bp was used to preprocess the data.

To obtain divergence estimates for each orthogroup, we used the branch model (model=2, nssites=0) in the CODEML package in PAML v4.7a (Yang 2007), using the specified phylogeny; ((*A. cygnoides*, *A. platyrhynchos*), (*N. meleagris*, (*P. cristatus*, (*M. gallopavo*, *P. colchicus*))), *T. guttata*). The branch model was used to calculate mean d_N/d_S across all Galloanserae branches, excluding the *T. guttata* outgroup. We will refer to this as the Galloanserae analysis. We also used the branch model to calculate mean d_N/d_S for each of the six Galloanserae species separately. Specifically, for each species, we calculated mean d_N/d_S from the terminal tip to the Galloanserae common ancestor. We will refer to this as the species-specific analysis. This approach ensures that the branch length over which d_N/d_S is calculated is identical for each species and therefore prevents interspecific variation in branch length biasing our conclusions (Montgomery *et al.* 2011). As mutational saturation and double hits can lead to inaccurate divergence estimates (Axelsson *et al.* 2008), orthogroups were excluded if tree length d_S >2 across all branches.

Using sequence divergence to estimate the Faster-Z Effect

The avian genome exhibits considerable karyotypic variation in chromosome size. Therefore, mean d_N, d_S and d_N/d_S were calculated separately for all autosomes, autosomes 1–10, microchromosomes and the Z chromosome. Microchromosomes exhibit an elevated recombination rate, greater gene density and GC content, all of which have been shown to impact the nature and efficacy of selection (Burt 2002; Ellegren 2013). The fairest measure of Faster-Z Evolution is therefore to contrast divergence between the Z chromosome and similar-sized autosomes 1–10 (Mank *et al.* 2010b).

For each genomic category, mean d_N and mean d_S were calculated as the sum of the number of substitutions across all contigs in a given category divided by the number of sites (d_N = sum D_N/sum N, d_S = sum D_S/sum S, where $D_{N/S}$ is an estimate of the number of nonsynonymous/synonymous substitutions and N/S is the number of nonsynonymous/synonymous sites). This approach avoids the problems of infinitely high d_N/d_S estimates arising from contigs with extremely low d_S (Mank *et al.* 2007a, 2010b) and prevents disproportionate weighting of shorter contigs.

Bootstrapping with 1000 repetitions was used to generate 95% confidence intervals, and significant differences between genomic categories were determined from 1000 permutation tests. One-tailed *P*-values are reported because we specifically test whether d_N, d_S and d_N/d_S are significantly higher for Z-linked contigs vs. autosomal contigs. Mean Z-linked and autosomal d_N, d_S and d_N/d_S values were calculated for the whole Galloanserae (Galloanserae analysis) and for each of the six species (species-specific analysis). Faster-Z Effect was calculated as d_{NZ}/d_{SZ}: d_{NA}/d_{SA}.

Testing the relationship between sexual selection and Faster-Z Effect

To test the hypothesis that the magnitude of Faster-Z increases with increased variance in male reproductive success, we performed phylogenetically controlled regression analyses between Faster-Z (d_{NZ}/d_{SZ}: d_{NA}/d_{SA}) and relative N_{EZ} for each Galloanserae species and two measures of female promiscuity. The intensity of sperm competition, a widely used proxy for the magnitude of postcopulatory sexual selection and therefore variance in male reproductive success, is strongly predicted by relative testes weight and sperm number (Moller 1991; Moller & Briskie 1995; Birkhead & Moller 1998). These measures are also frequently used to test genotype–phenotype hypotheses (e.g. Dorus *et al.* 2004; Ramm *et al.* 2008). Residual testes weight was calculated using the following equation describing the linear relationship between log testes weight and body weight across a large number of birds (Pitcher *et al.* 2005): $\log_2$[testes mass(g)] = −1.56 + 0.61 $\log_2$ [body mass(g)] (Moller 1988, 1991; Birkhead & Petrie 1995). For all six species in this study, relative testes weight was less than expected given body weight. Log sperm number (10^6) has been measured in previous studies (Moller 1988, 1991; Birkhead & Petrie 1995). Estimates for body weight and sperm number were not available for *A. cygnoides* and therefore *A. anser* estimates were used instead, as these species are closely related (Ruokonen *et al.* 2000) and both exhibit strictly monogamous mating systems.

These analyses were performed using phylogenetic generalized least squares models (PGLS) in BAYESTRAITS v2-beta (Pagel 1999; Pagel *et al.* 2004) with maximum likelihood and 1000 runs for each analysis. PGLS corrects for phylogenetic nonindependence. Phylogenies were obtained from birdtree.org using the Ericson data set. For each regression analysis, mean r^2 and mean *t*-value (mean regression coefficient/mean standard error) were calculated. A one-tailed *t*-test with four degrees of freedom was used to determine whether the slope was significantly >0.

Differences in the rate of male-biased mutation across the six species could contribute to variation in Faster-Z Effect because the Z chromosome is more often present in males than the autosomes (Kirkpatrick & Hall 2004a). We explicitly tested for significant differences in mean Z-linked d_S across the six species using permutation tests with 1000 replicates to verify that were no underlying differences in mutation rate.

Tests of positive selection using sequence data

To test for signatures of positive selection acting at a subset of sites, we used the site models in the CODEML package in PAML v4.7a (Yang 2007). These models allow d_N/d_S to vary among sites but not across lineages. To test for positive selection, we compared likelihoods from two models; M1a (Nearly neutral, model=0, nssites=1) and M2a (Positive selection, model=0, nssites=2). Under model M1a, sites can fall into one of two categories (purifying selection d_N/d_S <1 and neutral evolution d_N/d_S = 1), whereas there is an additional category under model M2a (positive selection d_N/d_S >1).

The following phylogeny was specified; ((*A. cygnoides*, *A. platyrhynchos*), (*N. meleagris*, (*P. cristatus*, (*M. gallopavo*, *P. colchicus*))), *T. guttata*).

Tests of positive selection using polymorphism data

We tested for deviations from neutrality using polymorphism data. Polymorphism data was obtained by first mapping RNA-seq reads to orthogroups using the two-pass alignment method of the STAR aligner with default parameters (Dobin *et al.* 2013). SNPs were called using VARSCAN v2.3.6 (Koboldt *et al.* 2009, 2012) and SAMTOOLS (Li *et al.* 2009) following the recommendations of Quinn *et al.* 2013 (Quinn *et al.* 2013). Only uniquely mapping reads were used to call SNPs. SAMTOOLS was run with probabilistic alignment disabled and a maximum read depth of 10 000 000. VARSCAN mpileup2snp was run with a minimum coverage of 2, a minimum average quality of 20, with the strand filter, *P*-value of 1, a minimum variant allele frequency threshold of 1E-1 and a minimum frequency to call homozygote of 0.85. SNPs were required to have a minor allele frequency >0.15 and to be from regions where at least 4 samples had a read depth >20 and have a Phred quality >20. Valid SNPs were matched to the reading frame to determine whether they were synonymous or nonsynonymous. Fixed sites were identified using the same quality and coverage thresholds used to call SNPs.

We explicitly tested whether our power to identify SNPs is equal across the Z and autosomes, despite differences in sequencing coverage. We generated random diploid populations of individuals with varying minor allele frequencies. From these populations, we sampled 20 (autosomal) and 15 (Z-linked) alleles separately 1000 times without replacement and for each sample determined the presence or absence of polymorphism. At a minor allele frequency of 0.15%, the false-negative rate for both the autosomes and Z chromosome was very low (autosomes = 0.023, Z chromosome = 0.068), although marginally lower for the autosomes. We also repeated analyses using a minor allele frequency threshold of 25% (false-negative rate autosomes = 0.001, Z chromosome = 0.009); however, our power is limited at this threshold due to a large reduction in detectable SNPs (Tables S3 and S4, Supporting Information). Our conclusions were broadly comparable across both minor allele frequency thresholds.

For each species, mean nonsynonymous polymorphism (p_N), synonymous polymorphism (p_S) and p_N/p_S were calculated separately for Z-linked and autosomal 1–10 orthogroups. Specifically, mean polymorphism was calculated as the sum of the number of polymorphic sites across all contigs in a given genomic category divided by the number of sites (p_N = sum P_N/sum N, p_S = sum P_S/sum S where $P_{N/S}$ is the number of nonsynonymous/synonymous polymorphic sites and N/S is the number of nonsynonymous/synonymous sites). Faster-Z was calculated as p_{NZ}/p_{SZ}: p_{NA}/p_{SA}. Bootstrapping with 1000 repetitions was used to generate 95% confidence intervals, and significance differences between genomic categories were determined from 1000 permutation tests.

For each species, we used the McDonald–Kreitman test (McDonald & Kreitman 1991) to estimate the number of contigs evolving under adaptive and neutral evolution. The McDonald–Kreitman test contrasts the number of nonsynonymous and synonymous substitutions (D_N and D_S) with polymorphisms (P_N and P_S). D_N and D_S for each species were obtained from the species-specific PAML analysis, where divergence was calculated from the terminal tip to the Galloanserae common ancestor, excluding the *T. guttata* outgroup. A deficit of nonsynonymous polymorphisms relative to substitutions is indicative of positive selection [(D_N/D_S) > (P_N/P_S)], and an excess of nonsynonymous polymorphisms relative to substitutions is indicative of relaxed purifying selection [(D_N/D_S) < (P_N/P_S)]. For each contig, we tested for departures from neutrality using a 2 × 2 contingency table and Pearson's chi-squared test (Hope 1968; Patefield 1981) in R v3.1.0 (R Core Team 2014). Contigs were only included in the analysis if the sum of each marginal row and column of the 2 × 2 contingency table was greater or equal than 6 (Begun *et al.* 2007; Andolfatto 2008). We used the qvalue function in R with a false discovery rate = 0.05 and lambda = 0 to correct for multiple testing. After identifying contigs with signatures of positive selection, we tested for significant differences in the proportion of these contigs on the Z chromosome vs. the autosomes using Pearson's chi-squared test in R.

Lastly, we used polymorphism data to test for an excess or under-representation of Z-linked nonsynonymous polymorphisms relative to the autosomes. Excess or underrepresentation is indicative of relaxed purifying selection or positive selection, respectively. For this analysis, we separately concatenated P_N and P_S for each species and used Pearson's chi-squared test to test for significant differences in P_N/P_S between the Z chromosome and autosomes (Mank *et al.* 2007a).

Calculating relative effective population size of the Z chromosome

We calculated the effective population size (N_E) of the Z chromosome and autosomes 1–10 for each species using two separate approaches based on π and θ.

For each contig, the number of fourfold degenerate sites (4D) and polymorphic fourfold degenerate sites

(P_{4D}) was calculated. Nucleotide diversity was calculated for each genomic category as π = sum P_{4D}/sum 4D. Watterson's estimator of theta (θ) (Watterson 1975) was also calculated as θ = sum 4D/(sum[i = 1…n−1] 1/i) where n is the number of chromosomes in the sample. θ per site was then calculated. Finally, we recalculated π and θ using all polymorphic synonymous sites.

Effective population size was calculated separately for the Z and autosomes as N_E = (π or θ)/[4*(U*generation time)]. The mutation rate per site per year (U) was calculated separately for the Z chromosome (1.45E-09) and autosomes (1.33E-09) to account for male-mutation bias, using previous Galliform estimates of Z-linked and autosomal divergence (Dimcheff *et al.* 2002; Axelsson *et al.* 2004; van Tuinen & Dyke 2004; Mank *et al.* 2010a). U = K/2T, where K is the no of substitutions per site between homologous sequences and T is divergence time. Bootstrapping with 1000 repetitions was used to generate 95% confidence intervals for effective population size estimates.

Tests of positive selection using gene expression

The relative role of selection vs. drift in driving Faster-Z Evolution can be disentangled using gene expression (Baines *et al.* 2008; Mank *et al.* 2010b; Sackton *et al.* 2014). Gene expression was quantified using only adult gonad samples, because this tissue exhibits the greatest magnitude of sex-biased transcription (Mank *et al.* 2007b; Pointer *et al.* 2013) and therefore maximizes the number of female-biased contigs used in the analysis. Expression was estimated as reads per kilobase per million mappable reads (RPKM) and normalized to control for differences in sequencing depth across samples (Brawand *et al.* 2011).

Mean male and female RPKM of each orthogroup were calculated separately for each species, together with fold change [a measure of sex-bias: $\log_2$(male RPKM)-$\log_2$(female RPKM)]. A *t*-test was used to identify significantly sex-biased contigs, and the Benjamini–Hochberg method (FDR of 5%) (Benjamini & Hochberg 1995) used to correct for multiple testing (Mank *et al.* 2010c; Pointer *et al.* 2013; Perry *et al.* 2014). Female-biased and male-biased contigs were classified as significantly sex-biased ($P < 0.05$) or sex-limited with a $\log_2$ fold change of <−1 and >1, respectively. Unbiased contigs had a $\log_2$ fold change between <1 and >−1.

To verify that our method of defining sex bias was consistent with other approaches, we also used EDGER to categorize sex bias and compared the overlap between both approaches. Briefly, for each species, we extracted raw read counts for 2 RPKM filtered contigs from RSEM (Li & Dewey 2011), normalized to control for differences in sequencing depth across samples using TMM in EDGER and tested for sex-biased gene expression using the exactTest function in EDGER (Robinson & Oshlack 2010; Robinson *et al.* 2010; McCarthy *et al.* 2012). Female-biased and male-biased contigs were classified as above using a significant *P*-value and $\log_2$ fold change of <−1 and >1, respectively. Our approach of categorizing sex bias was consistent with the results from EDGER, and we observe an overlap of 89–96% between expression categories as defined by both approaches.

We used three approaches to test the predictions of the selection and drift hypotheses. First, we calculated Faster-Z for orthogroups where expression category was conserved across all six species. This was to avoid diluting significant signals of selection or drift by including orthogroups where exposure to the dominant evolutionary force has not been consistent over time due to rapid expression turnover. Mean d_N, d_S and d_N/d_S were calculated separately for each expression category for Z-linked and autosomal contigs using divergence estimates from the Galloanserae analysis in CODEML (Yang 2007). Bootstrapping with 1000 repetitions was used to generate 95% confidence intervals. Significant differences between genomic categories were determined using permutation tests with 1000 repetitions.

We then repeated this analysis with relaxed criteria to maximize the number of orthogroups in each expression category. Specifically, we compared the Faster-Z Effect between putatively female-biased contigs (defined as contigs where at least half of the species had female-limited or significantly female-biased expression, and the fold change was <0 across all species) and male-biased contigs (where at least half of the species had male-limited or significantly male-biased expression, and the fold change was >0 across all species).

Finally, we assessed the relationship between species-specific Faster-Z Evolution and gene expression. For each species, we separately calculated d_{NZ}/d_{SZ}: d_{NA}/d_{SA} for female-, male- and unbiased contigs for each species as defined with *t*-tests and fold change thresholds. Significance was assessed using permutation tests with 1000 repetitions.

Gene ontology analysis

We used GORILLA (Eden *et al.* 2007, 2009) to perform a Gene Ontology enrichment analysis to test for enriched gene function terms for Z-linked contigs compared with the autosomes. Mouse reciprocal orthologs were identified using BIOMART (ENSEMBL v.77) for Z-linked and autosomal 1–10 orthologs. The target list contained Z-linked orthologs and the background list contained autosomal orthologs. *P*-values were corrected for multiple testing using the Benjamini–Hochberg method (Benjamini & Hochberg 1995).

Results

Faster–Z Evolution

We assembled *de novo* transcriptomes for six Galloanserae species, spanning approximately 90 million years of avian evolution van Tuinen and Hedges (2001) (Fig. 2), and identified 160 Z-linked and 2431 autosomal orthogroups. Across the Galloanserae, mean d_N/d_S of the Z chromosome is significantly higher than that of the autosomes, due to significantly elevated d_{NZ} (Table 1, Fig. 3). There is no difference in d_S between the Z chromosomes and all autosomes ($P = 0.865$).

Seven-hundred and forty-one autosomal orthogroups are located on microchromosomes in the chicken genome, and microchromosomes exhibit different genomic properties to the rest of the autosomes. These properties impact the nature and efficacy of selection (Burt 2002; Ellegren 2013); therefore, the fairest measure of Faster-Z Evolution is to contrast divergence between the Z chromosome and similar-sized autosomes 1–10 (Mank *et al.* 2010b). We identified 1690 orthogroups located on autosomes 1–10. Mean d_{NZ}/d_{SZ} and d_{NZ} are both significantly higher than mean d_N/d_S and d_N of autosomal 1–10 orthogroups (Table 1, Fig. 3). This pattern is consistent with the results of the previous analysis using all autosomes, and with previous estimates of Faster-Z Evolution in birds (Mank *et al.* 2007a, 2010b; Dalloul *et al.* 2010; Ellegren *et al.* 2012; Wang *et al.* 2014a). For the rest of the manuscript, autosomal will refer to autosomal 1–10 orthogroups and Faster-Z will refer to the comparison between Z-linked and autosomal 1–10 orthogroups d_{NZ}/d_{SZ}: d_{NA}/d_{SA}.

In each of the six Galloanserae species, d_{NZ}/d_{SZ} is higher than d_{NA}/d_{SA} based on the species-specific analysis, and there is interspecific variation in the magnitude of this difference (Table 2). We find no significant difference in d_S between the Z chromosome and autosomes for any species, consistent with previous findings that male-biased mutation rate is weak across the Galloanserae (Bartosch-Harlid *et al.* 2003; Axelsson *et al.* 2004). This suggests that Z-linked mutation rate does not vary significantly across the six species (addressed further in the Discussion).

Variation in sperm competition drives Faster-Z Evolution

The intensity of sperm competition, a widely used indicator of postcopulatory sexual selection and therefore one measure of variance in male mating success, is strongly predicted by relative testes weight and sperm number in birds (Moller 1991; Birkhead & Moller 1998;

Table 1 d_N, d_S and d_N/d_S for Z-linked and autosomal genes across Galloanserae clade

	Z chromosome (160 contigs)	Autosomes 1–10 (1690 contigs)	Microchromosomes (741 contigs)	All autosomes (2431 contigs)
d_S 95% CI	0.432 (0.413–0.454)	0.424 (0.417–0.432) $P = 0.229$	0.510 (0.493–0.528) $P = 1.000$	0.447 (0.440–0.454) $P = 0.865$
d_N 95% CI	0.056 (0.049–0.065)	**0.047 (0.044–0.049)** $\mathbf{P = 0.007}$	**0.040 (0.037–0.043)** $\mathbf{P < 0.001}$	**0.045 (0.042–0.047)** $\mathbf{P = 0.005}$

Significance values were determined from 1000 permutation tests, and bootstrapping with 1000 repetitions was used to generate 95% confidence intervals. Significant differences between autosomal and Z-linked orthogroups are in bold.

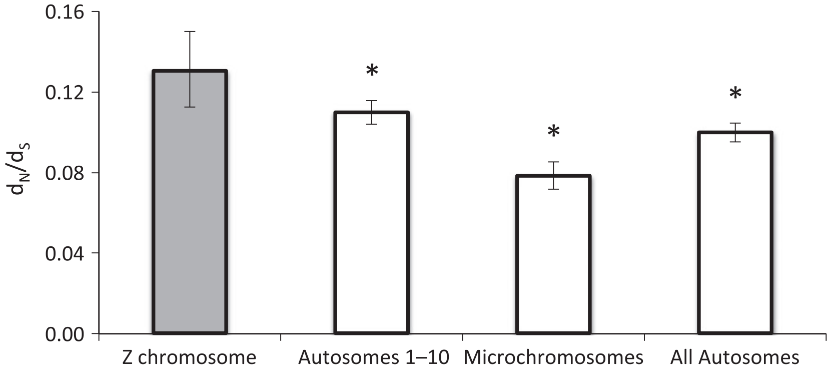


Fig. 3 Estimates of mean d_N/d_S for loci on autosomes and the Z chromosome across the Galloanserae. Synonymous and nonsynonymous divergence estimates were calculated using the branch model in PAML (Galloanserae analysis). 95% confidence intervals were calculated by bootstrapping with 1000 replicates, and significant differences in d_N/d_S between autosomal and Z-linked orthogroups (permutation test, 1000 replicates) are indicated (*).

Table 2 d_N, d_S and d_N/d_S for Z-linked and autosomal genes across Galloanserae species

	Z chromosome			Autosomes 1–10			Faster-Z Effect
Species	d_N (95% CI)	d_S (95% CI)	d_N/d_S (95% CI)	d_N (95% CI)	d_S (95% CI)	d_N/d_S (95% CI)	d_{NZ}/d_{SZ}: d_{NA}/d_{SA} (95% CI)
Meleagris gallopavo	0.023 (0.020–0.027)	0.163 (0.155–0.170)	0.144 (0.123–0.165)	**0.019** **(0.018–0.020)** **P = 0.005**	0.158 (0.154–0.161) P = 0.168	**0.120** **(0.113–0.127)** **P = 0.011**	1.205 (1.035–1.390)
Phasianus colchicus	0.021 (0.018–0.025)	0.161 (0.153–0.168)	0.134 (0.114–0.154)	**0.018** **(0.017–0.020)** **P = 0.035**	0.157 (0.153–0.160) P = 0.215	0.118 (0.111–0.125) P = 0.061	1.137 (0.961–1.331)
Numida meleagris	0.019 (0.016–0.022)	0.133 (0.127–0.140)	0.140 (0.119–0.162)	**0.016** **(0.015–0.017)** **P = 0.041**	0.132 (0.129–0.135) P = 0.393	**0.123** **(0.116–0.130)** **P = 0.049**	1.140 (0.965–1.332)
Anas platyrhynchos	0.015 (0.012–0.018)	0.116 (0.108–0.126)	0.131 (0.107–0.155)	**0.013** **(0.012–0.014)** **P = 0.030**	0.116 (0.113–0.119) P = 0.518	**0.109** **(0.103–0.116)** **P = 0.024**	1.200 (0.974–1.457)
Anser cygnoides	0.012 (0.010–0.015)	0.100 (0.093–0.107)	0.125 (0.103–0.148)	0.011 (0.010–0.012) P = 0.083	0.099 (0.097–0.101) P = 0.378	0.111 (0.105–0.118) P = 0.083	1.129 (0.939–1.360)
Pavo cristatus	0.020 (0.017–0.023)	0.147 (0.139–0.154)	0.134 (0.114–0.157)	0.017 (0.016–0.018) P = 0.068	0.147 (0.144–0.150) P = 0.502	0.118 (0.112–0.125) P = 0.056	1.133 (0.951–1.303)

Significance values were determined from 1000 permutation tests and bootstrapping with 1000 repetitions was used to generate 95% confidence intervals. Significant differences between autosomal and Z-linked orthologs are shown in bold.

Pitcher *et al.* 2005). We recovered a significant positive association between magnitude of Faster-Z Evolution and both log sperm number (r^2 = 0.684, P = 0.011, t_4 = 3.629) and residual testes weight (r^2 = 0.552, P = 0.026, t_4 = 2.744) after correcting for phylogeny (Fig. 4). To test the strength of these associations, we sequentially removed each species and repeated the analyses (Table S5). Despite the reduction in sample size and therefore statistical power, there was no change to the significance or direction of the slope for log sperm number. For residual testes weight, there was no change to the direction of the slope but when either *A. cygnoides* or *A. platyrhynchos* was excluded, the relationship was nonsignificant (Table S5).

There are two plausible explanations for our finding that the magnitude of Z-linked divergence increases with increasing female promiscuity. A recent study in silk moths has shown that Faster-Z Evolution is adaptive, and results from increased efficacy of selection acting on recessive advantageous mutations in the hemizygous sex (Sackton *et al.* 2014). Conversely, a study in birds suggested that avian Faster-Z Evolution

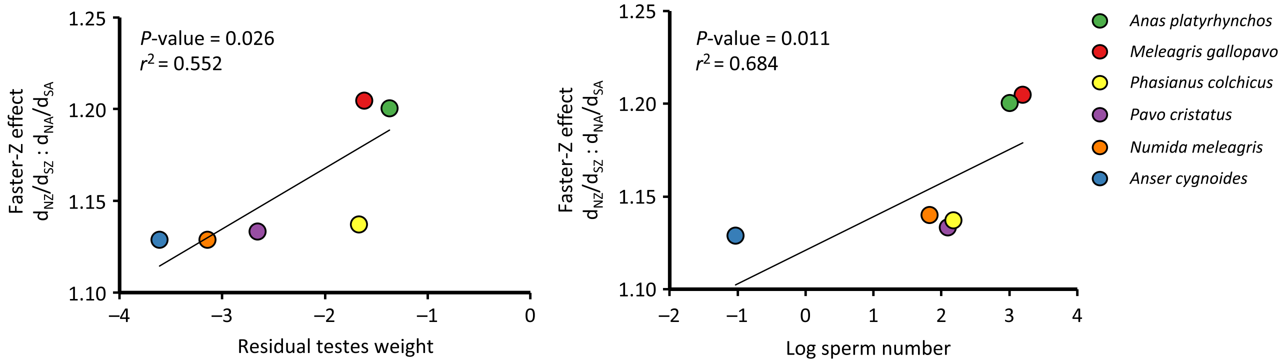


Fig. 4 Phylogenetically controlled regression between proxies of sperm competition and Faster-Z Effect. Data points are raw species values but P-values and r^2 estimates were calculated using phylogenetic generalized least squares regression with maximum likelihood and 1000 runs for each analysis. Autosomes refers to macrochromosomes (autosomes 1–10).

Table 3 Effective population size estimates of the Z chromosome and autosomes

Species	N_{EZ} (E + 05) (95% CI)	$N_{EA1–10}$ (E + 05) (95% CI)	$N_{EZ}/N_{EA1–10}$ (95% CI)
Meleagris gallopavo	1.761 (1.087–2.702)	6.047 (5.656–6.469)	0.291 (0.179–0.426)
Phasianus colchicus	3.188 (2.308–4.210)	9.481 (8.948–10.054)	0.336 (0.234–0.460)
Numida meleagris	1.695 (0.773–3.213)	7.233 (6.682–7.848)	0.234 (0.103–0.423)
Anas platyrhynchos	6.150 (3.927–8.758)	18.427 (17.447–19.544)	0.334 (0.209–0.470)
Anser cygnoides	4.045 (2.774–5.591)	10.894 (10.233–11.570)	0.371 (0.250–0.529)
Pavo cristatus	1.088 (0.167–2.811)	2.393 (2.095–2.697)	0.455 (0.057–1.227)

N_E was calculated using the same method as Mank *et al.* 2010b;. Mutation rate estimates are from Axelsson *et al.* 2004; Dimcheff *et al.* 2002 and van Tuinen & Dyke 2004.
Minor allele frequency threshold of 0.15.
Nucleotide diversity (π) was calculating using fourfold degenerate sites.

is a neutral process, driven by relaxed efficacy of purifying selection as a consequence of relative differences in N_{EZ}/N_{EA} (Mank *et al.* 2010b). Under the latter hypothesis, variation in male reproductive success, associated with sexual selection, is predicted to alter the relationship between N_{EZ} and N_{EA}, and therefore the relative magnitude of drift acting on the Z chromosome (Charlesworth *et al.* 1993; Vicoso & Charlesworth 2009). Specifically, with increasing variance in male reproductive success, relative N_{EZ} decreases, resulting in greater magnitude of drift and therefore Faster-Z Effect (Wright & Mank 2013).

We use sequence divergence, polymorphism and expression data to test whether the relationship between female promiscuity and Faster-Z Evolution is adaptive or neutral.

Estimates of relative N_{EZ}

After filtering for quality and read depth, across Z-linked and autosomal 1–10 contigs, we identified 12 436 SNPs in *A. platyrhynchos,* 4584 in *M. gallopavo,* 6850 in *P. colchicus,* 5205 in *N. meleagris,* 2012 in *P. cristatus* and 8128 in *A. cygnoides* (Table S3).

For each species, we calculated the effective population size of the Z chromosome (N_{EZ}) and autosomes 1–10 (N_{EA}) using a number of approaches. We accounted for male-biased mutation rate and generation time using previous Galliform estimates (Dimcheff *et al.* 2002; Axelsson *et al.* 2004; van Tuinen & Dyke 2004; Mank *et al.* 2010a) (Table 3, Tables S6, S7 and S8, Supporting Information) (Vicoso & Charlesworth 2009). Under strict monogamy, N_{EZ} is predicted to equal ¾ N_{EA}. For all species with the exception of *P. cristatus,* N_{EZ} was significantly <¾ N_{EA}. However, the 95% CI for this species was unusually wide, probably as a result of the low frequency of SNPs detected (Table S3).

The relationship between N_{EZ}/N_{EA} and sperm number, residual testes weight or Faster-Z was not statistically significant (sperm number: $r^2 = 0.083$, $P = 0.252$, $t_4 = 0.735$; residual testes weight: $r^2 = 0.068$, $P = 0.275$, $t_4 = 0.656$; Faster-Z: $r^2 = 0.220$, $P = 0.132$, $t_4 = 1.300$; Table S9, Supporting Information). Additionally, the autosomal effective population size of *P. cristatus* is significantly smaller than the other six species, indicating either a very recent bottleneck or variation in family structure across the individuals sampled in this study. This finding hints at the sensitivity of N_E calculations to many factors (Hartl & Clark 2007), including recombination rate and recent demographic perturbations (Pool & Nielsen 2007). This may explain both the unusually low N_E estimates in *P. cristatus* as well as the lack of significant association between N_{EZ}/N_{EA} and measures of sperm competition (addressed further in the Discussion).

Tests of positive selection

We used sequence and polymorphism data from our six species to test whether selection is more effective for Z-linked vs. autosomal loci. Using the site-model test in CODEML, we found significant evidence for positive selection acting on 5/160 Z-linked loci (1/160 after sequential Bonferroni's correction) and 51/1690 autosomal loci (5/1690 after sequential Bonferroni's correction) (Table 4, Table S10, Supporting Information). There was no significant difference in the proportion of positively selected loci on the Z chromosome or autosomes 1–10 either before or after multiple testing correction (χ^2, d.f. = 1, $P > 0.400$ in both comparisons). This indicates that selection is not more effective on the Z chromosome; however, the power of this analysis is limited by the low number of total contigs under positive selection.

We next used polymorphism data to test for deviations from neutrality. With the exception of *N. meleagris* and *P. cristatus*, p_{NZ}/p_{SZ} is significantly greater than p_{NA}/p_{SA} (Table 5, Table S11, Supporting Information). This finding of excess nonsynonymous polymorphism

Table 4 Site-model test results for contigs under positive selection

G. gallus ortholog*	Chromosome	ω	Proportion of sites	M1a likelihood ratio	M2a likelihood ratio	LRT	*P*-value	*P*-fdr value†
22552	1	2.897	0.122	−6535.857	−6522.227	27.259	<0.001	0.003
21101	1	4.155	0.033	−14063.297	−14050.286	26.023	<0.001	0.006
31776	3	4.608	0.130	−1270.098	−1256.430	27.337	<0.001	0.003
39919	6	4.226	0.310	−1630.735	−1611.278	38.915	<0.001	<0.001
03831	8	4.817	0.080	−9607.226	−9560.287	93.878	<0.001	<0.001
10504	15	3.343	0.072	−5389.616	−5375.473	28.287	<0.001	0.002
01868	20	9.422	0.013	−4192.958	−4179.195	27.526	<0.001	0.003
02022	28	4.914	0.068	−2768.690	−2753.634	30.110	<0.001	0.001

*ENSGALT000000.
†Sequential Bonferroni's correction (Holm 1979).

on the Z chromosome relative to the autosomes suggests that selection is less effective at removing mildly deleterious mutations from the Z chromosome. This finding is consistent with the drift hypothesis of Faster-Z, rather than the adaptive hypothesis. Interestingly, *N. meleagris* Z chromosome exhibits a nonsignificant deficit of p_N, potentially as a consequence of monogamy, which would maximize N_{EZ}/N_{EA} and therefore the potential of selection to act on the Z chromosome in this species.

For each species, we estimated the number of contigs evolving under adaptive evolution using the McDonald–Kreitman test (McDonald & Kreitman 1991). This test contrasts the number of nonsynonymous and synonymous substitutions (D_N and D_S) with polymorphisms (P_N and P_S) for each contig. An excess of nonsynonymous substitutions relative to polymorphism is indicative of positive selection [$(D_N/D_S) > (P_N/P_S)$], and under-representation of nonsynonymous substitutions relative to polymorphism is indicative of relaxed purifying selection [$(D_N/D_S) < (P_N/P_S)$]. We detected no Z-linked contigs with signatures of positive selection, and there was no difference between the Z chromosome and autosomes 1–10 in the proportion of loci under positive selection in any species (χ^2, d.f. = 1, $P > 0.500$ in all cases) (Table S12, Supporting Information). However, only contigs with sufficient numbers of substitutions and polymorphisms were included in the analysis (Begun *et al.* 2007; Andolfatto 2008), and therefore, our ability to draw species-specific conclusions is limited by low sample sizes.

Lastly, for each species, we concatenated the number of P_N and P_S across all Z-linked and all autosomal 1–10 contigs separately (Table 6, Table S13, Supporting Information) and tested for significant differences between Z-linked and autosomal P_N/P_S. For each species, there is a significant excess of Z-linked nonsynonymous polymorphism relative to the autosomes for all species with the exceptions of *P. cristatus* and *N. meleagris*. This is again consistent with a reduction in the power of selection to remove mildly deleterious alleles from this chromosome.

The lack of difference in Z-linked and autosomal nonsynonymous polymorphism in *P. cristatus* and *N. meleagris* could be attributed to a number of factors. It could reflect biological differences in sexual selection and therefore the magnitude of drift acting on the Z chromosome. However, although this explanation is consistent with the monogamous mating system of *N. meleagris*, it is not consistent with the *P. cristatus*, which exhibits a lek mating system (Petrie *et al.* 1999). More likely, this pattern reflects the limitations of polymorphism data and the difficulty in controlling for family structure and demographic effects (Hartl & Clark 2007). For example, the number of SNPs in *P. cristatus* is much lower than the other five species, and therefore, the statistical power of this analysis is limited (Table 6).

Differences in gene content between the sex chromosomes and autosomes can contribute to observed patterns of Faster-Z/X (Meisel & Connallon 2013) by biasing the potential for positive selection in different genomic categories. However, the results of our GORILLA functional enrichment test reveal no significantly enriched gene ontology terms for Z-linked orthogroups compared with autosomes 1–10 after correcting for multiple tests.

Gene expression

We used gene expression data from gonads of our six avian species to identify the dominant force driving Faster-Z Evolution across the Galloanserae clade. If Faster-Z Evolution is adaptive and driven by increased efficacy of selection acting on recessive mutations in the hemizygous sex, we predict the Faster-Z Effect to be largest for female-biased, followed by unbiased and then male-biased genes. If it is due to neutral causes, there will be no difference in the rate of Faster-Z

Table 5 p_N, p_S and p_N/p_S for Z-linked and autosomal genes across Galloanserae species

	Z chromosome			Autosomes 1–10			Faster-Z Effect
Species	p_N (95% CI)	p_S (95% CI)	p_N/p_S (95% CI)	p_N (95% CI)	p_S (95% CI)	p_N/p_S (95% CI)	p_{NZ}/p_{SZ}: p_{NA}/p_{SA} (95% CI)
Meleagris gallopavo	0.000 (0.000–0.000)	0.001 (0.001–0.002)	0.176 (0.112–0.256)	0.000 (0.000–0.001) $P = 1.000$	0.005 (0.005–0.005) $P = 1.000$	**0.102 (0.093–0.111)** $\mathbf{P < 0.001}$	1.721 (1.039–2.716)
Phasianus colchicus	0.000 (0.000–0.001)	0.002 (0.002–0.003)	0.162 (0.109–0.236)	0.001 (0.001–0.001) $P = 1.000$	0.007 (0.007–0.008) $P = 1.000$	**0.095 (0.088–0.103)** $\mathbf{P < 0.001}$	1.704 (1.122–2.546)
Numida meleagris	0.000 (0.000–0.000)	0.002 (0.001–0.003)	0.083 (0.050–0.150)	0.001 (0.001–0.001) $P = 1.000$	0.005 (0.005–0.006) $P = 1.000$	0.102 (0.093–0.112) $P = 0.897$	0.813 (0.483–1.370)
Anas platyrhynchos	0.001 (0.000–0.001)	0.005 (0.004–0.008)	0.103 (0.064–0.156)	0.001 (0.001–0.001) $P = 1.000$	0.014 (0.013–0.015) $P = 1.000$	**0.072 (0.066–0.078)** $\mathbf{P = 0.002}$	1.426 (0.863–2.120)
Anser cygnoides	0.001 (0.000–0.001)	0.003 (0.002–0.004)	0.177 (0.109–0.262)	0.001 (0.001–0.001) $P = 0.999$	0.008 (0.008–0.009) $P = 1.000$	**0.108 (0.099–0.117)** $\mathbf{P < 0.001}$	1.642 (1.009–2.427)
Pavo cristatus	0.000 (0.000–0.000)	0.001 (0.000–0.002)	0.173 (0.096–0.541)	0.000 (0.000–0.000) $P = 0.930$	0.002 (0.002–0.002) $P = 1.000$	0.134 (0.116–0.156) $P = 0.137$	1.293 (0.681–4.002)

Significance values were determined from 1000 permutation tests, and bootstrapping with 1000 repetitions was used to generate 95% confidence intervals.
Significant differences between autosomal and Z-linked orthologs are shown in bold.
Minor allele frequency threshold of 0.15.

Evolution among expression classes (Baines *et al.* 2008; Mank *et al.* 2010b; Sackton *et al.* 2014). We tested this prediction at three levels in our data.

First, we identified orthogroups with consistent male-, female- and unbiased expression across all six species, thereby excluding any orthogroups where the nature of sex-bias, and therefore exposure to the dominant evolutionary force, has varied over Galloanserae evolutionary history. The rapid change in sex bias across this clade (Harrison *et al.* in press) means that relatively few genes are consistently sex-biased in our data set, resulting in 17 male-biased, 9 female-biased and 7 unbiased Z-linked orthogroups alongside 104 male-biased, 116 female-biased and 205 unbiased autosomal orthogroups. Among these gene sets, there was no significant difference in Faster-Z Effect (male-biased vs. female-biased $P = 0.542$, female-biased vs. unbiased $P = 1.000$, male-biased vs. unbiased $P = 0.616$, all two-tailed pairwise permutation tests with 1000 repetitions), shown in Fig. 5.

To exclude the possibility that we lack statistical power to distinguish between drift and selection due to low sample sizes, we next repeated the analysis and relaxed the definition of sex bias (see Materials and Methods). In doing so, we nearly doubled the number of orthogroups in each expression category; identifying 54 male-biased and 15 female-biased Z-linked orthogroups, together with 347 male-biased and 319 female-biased autosomal orthogroups. Again, there was no significant difference in Faster-Z Effect between these gene sets ($P = 0.916$, permutation test, 1000 repetitions), with female-biased d_{NZ}/d_{SZ}: $d_{NA}/d_{SA} = 1.491$ (95% CI = 0.997−2.137) and male-biased d_{NZ}/d_{SZ}: $d_{NA}/d_{SA} = 1.456$ (95% CI = 1.112−1.869).

Finally, we assessed whether there was any species-specific pattern in Faster-Z Evolution across male-, female- and unbiased contigs. There is no significant difference between Faster-Z of any expression category in any species after correction for multiple testing, with the exception of *N. meleagris* where we found a significantly larger Faster-Z Effect for male-biased compared with unbiased contigs (Tables S14 and S15, Supporting Information). At all three levels of analysis, our expression data are consistent with Faster-Z Evolution resulting predominantly from neutral forces.

Discussion

Faster rates of coding sequence evolution on the Z chromosome relative to the autosomes have been observed across a wide range of species (Mank *et al.* 2007a, 2010b; Dalloul *et al.* 2010; Ellegren *et al.* 2012; Sackton *et al.* 2014; Wang *et al.* 2014a,b); however, the underlying cause is unclear. Indirect evidence from an

Table 6 Significant differences between nonsynonymous and synonymous polymorphism on the Z chromosome and autosomes

	Z chromosome		Autosomes 1–10		
Species	P_N	P_S	P_N	P_S	Faster-Z Effect P_{NZ}/P_{SZ}: P_{NA}/P_{SA} *P*-value
Meleagris gallopavo	51	83	1174	3276	**1.715** ***P* = 0.004**
Phasianus colchicus	89	157	1654	4950	**1.700** ***P* < 0.001**
Numida meleagris	29	100	1339	3737	0.809 *P* = 0.372
Anas platyrhynchos	126	351	2417	9542	**1.417** ***P* = 0.001**
Anser cygnoides	127	206	2138	5657	**1.631** ***P* < 0.001**
Pavo cristatus	38	63	610	1301	1.286 *P* = 0.277

Significant differences were determined using Pearson's chi-squared test in R.
Significant differences between autosomal and Z-linked orthologs are shown in bold.
Minor allele frequency threshold of 0.15.

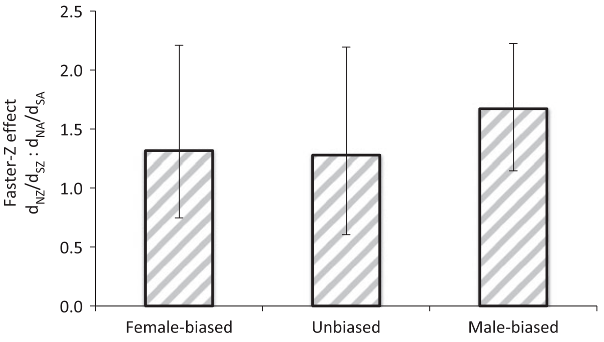


Fig. 5 Estimates of mean Faster-Z across sex-biased gene expression categories. Sex bias was defined using fold change thresholds and *t*-tests. 95% confidence intervals were calculated by bootstrapping with 1000 replicates. Autosomal orthologs were limited to chromosomes 1–10.

expression-based approach suggests that avian Faster-Z Evolution is driven by genetic drift (Mank *et al.* 2010b), but a recent study in silk moths postulated an adaptive explanation (Sackton *et al.* 2014). To determine the cause of Faster-Z Evolution in birds, we assembled *de novo* transcriptomes for six Galloanserae species, spanning 90 million years of avian evolution and combined expression, sequence and polymorphism data with measures of sperm competition and promiscuity. We present the first empirical evidence demonstrating the positive relationship between the Faster-Z Effect and measures of postcopulatory sexual selection and variance in male reproductive success.

This pattern is consistent with a considerable body of theory predicting that Faster-Z Evolution in birds is driven by changes in the relative strength of genetic drift as a result of increased variance in male reproductive success (Vicoso & Charlesworth 2009). In support of the predominant role of genetic drift in shaping rates of Z chromosome evolution, we used multiple sequence-, polymorphism- and expression-based approaches. Our expression analysis is consistent with previous work that found no difference in Faster-Z Evolution among sex-biased expression categories (Mank *et al.* 2010b). However, our analysis significantly extends this previous work by incorporating tests of positive selection based on divergence and polymorphism. The results from these multiple lines of evidence are broadly convergent, indicating that selection is not more effective on the Z chromosome. We conclude that Faster-Z Evolution in birds is due primarily to relaxed power of purifying selection and that the magnitude of this effect is dependent on the nature of sexual selection.

Promiscuity and sperm competition are drivers of Faster-Z Evolution

Changes in the skew of male reproductive success are commonly associated with promiscuity and the intensity of postcopulatory sexual selection (Andersson 1994), both of which decrease the N_{EZ}/N_{EA} ratio. If Faster-Z is neutral and nonadaptive, we predict that the magnitude of Faster-Z Evolution should increase as N_{EZ}/N_{EA} decreases (Vicoso & Charlesworth 2009), and therefore, we should expect both lower N_{EZ}/N_{EA} and increased rates of Faster-Z Evolution in promiscuous compared with monogamous populations (Fig. 1).

We uncovered a significant and positive association between the magnitude of Faster-Z and relative testes weight and sperm number, both reliable predictors of the intensity of sperm competition in birds (Fig. 4) (Moller 1991; Birkhead & Moller 1998; Pitcher *et al.* 2005). Sperm competition is a widely used indicator of the strength of postcopulatory sexual selection and therefore a good proxy for variance in male mating success and the magnitude of drift acting on the Z chromosome (Moller 1991; Birkhead & Moller 1998; Dorus *et al.* 2004). It is even possible we have underestimated the role of male mating success in driving Z chromosome divergence, as the birds sampled in this study have a lower testes weight than expected given their body weight (Pitcher *et al.* 2005).

Although the relationship between N_{EZ}/N_{EA} and sperm number or residual testes weight was not significant, N_{EZ}/N_{EA} across the Galloanserae is consistent with the nonadaptive hypothesis of Faster-Z Evolution (Vicoso & Charlesworth 2009) and is significantly less than the 0.75 predicted under strict monogamy, with the exception of *P. cristatus* (Table 3). We calculated effective population size using parameters estimated from previous Galliform studies (Dimcheff *et al.* 2002; Axelsson *et al.* 2004; van Tuinen & Dyke 2004; Mank *et al.* 2010a), and although mutation rate, male-biased mutation and generation time are not expected to vary substantially across the Galloanserae, we might expect slight differences. Overall N_E is also predicted to have a large effect on the magnitude of Faster-Z and relative contribution of nonadaptive and adaptive evolutionary forces. However, patterns of autosomal N_E do not reflect differences in Faster-Z across species.

Polymorphism estimates are sensitive to recent demographic perturbations, bottlenecks and recombination rate (Hartl & Clark 2007). Changes in population size have been shown to differentially impact N_{EZ} relative to N_{EA} and variation in population history across the Galloanserae may contribute to the lack of a significant relationship between N_{EZ}/N_{EA} and measures of promiscuity and sperm competition (Pool & Nielsen 2007). Previous attempts to estimate N_{EZ}/N_{EA} in birds (Corl & Ellegren 2012) showed sizable variation from what would be predicted by mating system, suggesting that N_{EZ}/N_{EA} estimates may simply be too inaccurate for the types of analyses used here. Because divergence data are not as sensitive to recent demographic perturbations, it can be argued that it is a fairer test for the role of male mating success and sperm competition in Faster-Z Evolution.

Tests of positive selection

We used sequence and polymorphism data to test the relative strength of selection on the Z chromosome vs. autosomes. In both the site-model tests in PAML as well as species-specific McDonald–Kreitman tests, there was no difference in the proportion of positively selected loci on the Z chromosome compared with the autosomes. The McDonald–Kreitman test is limited to sequences with sufficient numbers of substitutions and polymorphisms (McDonald & Kreitman 1991; Andolfatto 2008), and this restricted our analysis to a handful of Z-linked contigs. Therefore, to maximize the power of our data set, we concatenated polymorphism data across all Z-linked and autosomal contigs (Mank *et al.* 2007a). For the majority of species, an excess of Z-linked nonsynonymous polymorphism relative to the autosomes was observed, suggesting that selection is less able to purge mildly deleterious alleles from the Z chromosome. This pattern is consistent with the theoretical expectations of elevated levels of genetic drift. We would expect the opposite pattern, a deficit of Z-linked nonsynonymous polymorphism, under both positive and purifying selection.

Differences in gene content between the sex chromosomes and autosomes can bias the potential for positive selection to act on different genomic categories, and therefore may contribute to our observed patterns of Faster-Z (Meisel & Connallon 2013). The avian Z chromosome is enriched in male-biased genes (Mank & Ellegren 2009), which typically exhibit rapid rates of evolution (Meisel 2011; Parsch & Ellegren 2013). However, we do not find an elevated Faster-Z Effect for male-biased genes, and the results of our GORILLA functional enrichment analysis reinforce that differences in gene content are not likely to drive the pattern of Faster-Z we observe.

Overall, we failed to detect any indication that selection is more effective for Z-linked loci, consistent with the nonadaptive explanations for Faster-Z Evolution. However, it is important to note that our analyses are limited to orthologs conserved across 90 million years, and conservation across this span of time suggests that purifying selection is a dominant force acting on these genes. The important role of purifying selection in this gene set may bias our ability to detect positive selection using this data set. Nevertheless, our neutral explanation of Faster-Z is consistent with previous work indicating that sex chromosome dosage compensation status mediates the contribution of positive selection to Faster-Z Effect (Charlesworth *et al.* 1987; Mank 2009). Theory predicts that the adaptive component of Faster-Z is weaker in species with incomplete dosage compensation, such as birds (Ellegren *et al.* 2007; Mank 2009; Itoh *et al.* 2010; Uebbing *et al.* 2013), compared to those with complete dosage compensation.

Theory predicts that the magnitude of Faster-Z Effect should increase as N_{EZ}/N_{EA} decreases (Vicoso &

Charlesworth 2009), and therefore, we should expect increased rates of Faster-Z Evolution in promiscuous compared with monogamous populations. This prediction is consistent with our finding that Faster-Z is positively correlated with the intensity of sperm competition, and therefore variance in male reproductive success.

Faster-Z vs. Faster-X Evolution

Faster rates of coding sequence divergence have repeatedly been documented on the X and Z chromosomes relative to the autosomes, and there is considerable variation in the magnitude of this difference across species (Meisel & Connallon 2013). Moreover, there is a stark contrast between our results and those of Faster-X Evolution in *Drosophila* and mammals, where X-linked male-biased genes evolve more rapidly than unbiased and female-biased genes (Khaitovich *et al.* 2005; Baines *et al.* 2008; Grath & Parsch 2012). This pattern is consistent with an adaptive explanation of Faster-X Evolution driven by increased efficacy of selection acting on recessive mutations in the heterogametic sex. In addition, there is considerable evidence for signatures of adaptation on the X chromosome across many species (Thornton & Long 2005; Baines *et al.* 2008; Hvilsom *et al.* 2012; Langley *et al.* 2012).

The empirical evidence for neutral vs. adaptive explanations of Faster-Z and Faster-X Evolution, respectively, is supported by theoretical predictions (Vicoso & Charlesworth 2009). As variance in male reproductive fitness increases, $N_{EZ} < ¾ N_{EA}$, reducing the ability of selection to purge mildly deleterious alleles. In contrast, $N_{EX} > ¾ N_{EA}$ under increased variance in male reproductive success, indicating that Faster-X is more often due to positive selection acting on recessive mutations exposed in the heterogametic sex. However, a recent study in silk moths (Sackton *et al.* 2014) indicates that this prediction may not hold for all female heterogametic species and is dependent on numerous other factors, including overall population size and sex-specific recombination rates (Connallon *et al.* 2012).

Male-biased mutation

The relative rate of Z-linked divergence is thought to be influenced by multiple factors, not only variance in male reproductive success (Kirkpatrick & Hall 2004a; Connallon *et al.* 2012). The number of cell divisions, and therefore potential for mutations, is inherently higher in spermatogenesis compared with oogenesis. This male-biased mutation has been documented across a number of species (Bartosch-Harlid *et al.* 2003; Axelsson *et al.* 2004; Xu *et al.* 2012), and as the Z chromosome is present more often in males than females, it could contribute to the observed differences in relative Z-linked divergence (Kirkpatrick & Hall 2004a; Xu *et al.* 2012). However, previous estimates indicate the magnitude of male-biased mutation may be relatively weak across the Galloanserae (Bartosch-Harlid *et al.* 2003), ranging from 1.6 to 3.8 in Anseriformes (Wang *et al.* 2014b) and 1.7 to 2.52 in Galliformes (Axelsson *et al.* 2004). We failed to find a significant difference between d_{SZ} and d_{SA} in any species indicating that male-mutation bias does not vary significantly across this clade. This is consistent with the observation that the wild species in this study are seasonal breeders where spermatogenesis ceases in the nonbreeding season. Consequentially, the difference in number of meiotic cell divisions between males and females is reduced, and therefore, the potential for male-biased mutation is lower. In contrast, many previous estimates of male-biased mutation were based on domesticated species with continuous breeding cycles and spermatogenesis (Bartosch-Harlid *et al.* 2003; Axelsson *et al.* 2004). However, it is possible there is also a confounding effect of Z-linked codon usage bias, an excess of which has been observed on the *Drosophila* X chromosome (Singh *et al.* 2008).

Sexual selection and the Z chromosome

The sex chromosomes are predicted to play a disproportionate role in encoding sex-specific fitness due to their unequal inheritance pattern (Rice 1984). The Z chromosome in particular is thought to foster tight linkage between female preference genes and flashy male traits, and promote rapid evolution of some types of sexually selected traits (Rice 1984; Reeve & Pfennig 2003; Kirkpatrick & Hall 2004b). However, evidence that the Z chromosome harbours genes encoding sexually dimorphic phenotypes is mixed (Dean & Mank 2014). Z-linked male plumage genes have been documented in flycatchers (Saetre *et al.* 2003; Saether *et al.* 2007), but other studies have failed to find an association between sexually dimorphic traits and sex linkage (Knief *et al.* 2012; Schielzeth *et al.* 2012; Pointer *et al.* 2013). Our findings may help explain this discrepancy between theoretical and empirical data. The low effective population size of the Z chromosome relative to the autosomes may weaken the efficacy of sex-specific selection, particularly in the species under the strongest sexual selection regimes. This may limit the adaptive role of the Z chromosome in general, and in particular its role in encoding sexually selected traits. Given this, it is important to note that our results do not exclude the potential for selection acting on the Z chromosome, but suggests that relaxed purifying selection is more dominant on the Z chromosome relative to the autosomes.

Conclusions

We assessed the magnitude and drivers of Faster-Z Evolution across a clade of birds spanning 90 million years of evolution. Our analysis combines expression, sequence and polymorphism data with measures of sperm competition and promiscuity. The results from these multiple lines of evidence are broadly convergent, indicating that selection is less effective on the Z chromosome, and suggesting that Faster-Z Evolution in birds is due primarily to genetic drift. Moreover, we present the first empirical evidence demonstrating the positive relationship between the Faster-Z Effect and measures of promiscuity and sperm competition, and therefore variance in male mating success.

Acknowledgements

We thank Victoria Roberts and John Hall for collection assistance, and Geoff Parker for assistance in securing samples. We also thank Jen Perry, Rebecca Dean and Natasha Bloch for helpful suggestions and comments on the manuscript, Mario dos Reis for help with the PAML analyses and four reviewers for their extremely helpful comments. Sequencing was performed by the Wellcome Trust Centre for Human Genetics Sequencing Hub (funded by Wellcome Trust Grant reference 090532/Z09/Z and MRC Hub Grant G0900747 91070). The authors acknowledge the use of the UCL *Legion* High Performance Computing Facility (Legion@UCL) and UCL *Unity* SMP Facility, and associated support services, in the completion of this work. The work described here was funded by the European Research Council under the Framework 7 agreement (grant agreement 260233). A.E.W. is funded by a National Environment Research Council Ph.D. studentship. SHM is funded by a research fellowship from the Royal Commission for the Exhibition of 1851.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Andersson M (1994) *Sexual Selection*. Princeton University Press, New Jersey, USA.

Andolfatto P (2008) Controlling type-I error of the McDonald–Kreitman Test in genomewide scans for selection on noncoding DNA. *Genetics*, **180**, 1767–1771.

Axelsson E, Smith NGC, Sundstrom H, Berlin S, Ellegren H (2004) Male-biased mutation rate and divergence in autosomal, Z-linked and W-linked introns of chicken and turkey. *Molecular Biology and Evolution*, **21**, 1538–1547.

Axelsson E, Hultin-Rosenberg L, Brandström M, Zwahlen M, Clayton DF, Ellegren H (2008) Natural selection in avian protein-coding genes expressed in brain. *Molecular Ecology*, **17**, 3008–3017.

Baines JF, Sawyer SA, Hartl DL, Parsch J (2008) Effects of X-linkage and sex-biased gene expression on the rate of adaptive protein evolution in *Drosophila*. *Molecular Biology and Evolution*, **25**, 1639–1650.

Bartosch-Harlid A, Berlin S, Smith NGC, Moller AP, Ellegren H (2003) Life history and the male mutation bias. *Evolution*, **57**, 2398–2406.

Begun DJ, Holloway AK, Stevens K *et al.* (2007) Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biology*, **5**, 2534–2559.

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

Birkhead T, Moller AP (1998) *Sperm Competition and Sexual Selection*. Academic Press Inc., San Diego, California.

Birkhead TR, Petrie M (1995) Ejaculate features and sperm utilization in Peafowl *Pave Cristatus*. *Proceedings of the Royal Society B-Biological Sciences*, **261**, 153–158.

Brawand D, Soumillon M, Necsulea A *et al.* (2011) The evolution of gene expression levels in mammalian organs. *Nature*, **478**, 343–348.

Burt DW (2002) Origin and evolution of avian microchromosomes. *Cytogenetic and Genome Research*, **96**, 97–112.

Charlesworth B (2009) Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, **10**, 195–205.

Charlesworth B, Coyne JA, Barton NH (1987) The relative rates of evolution of sex chromosomes and autosomes. *The American Naturalist*, **130**, 113–146.

Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. *Genetics*, **134**, 1289–1303.

Connallon T, Singh ND, Clark AG (2012) Impact of genetic architecture on the relative rates of X versus autosomal adaptive substitution. *Molecular Biology and Evolution*, **29**, 1933–1942.

Corl A, Ellegren H (2012) The genomic signature of sexual selection in the genetic diversity of the sex chromosomes and autosomes. *Evolution*, **66**, 2138–2149.

Dalloul RA, Long JA, Zimin AV *et al.* (2010) Multi-platform next-generation sequencing of the domestic turkey (*Meleagris gallopavo*): genome assembly and analysis. *PLoS Biology*, **8**, e1000475.

Dean R, Mank JE (2014) The role of sex chromosomes in sexual dimorphism: discordance between molecular and phenotypic data. *Journal of Evolutionary Biology*, **27**, 1443–1453.

Dimcheff DE, Drovetski SV, Mindell DP (2002) Phylogeny of *Tetraoninae* and other galliform birds using mitochondrial 12S and ND2 genes. *Molecular Phylogenetics and Evolution*, **24**, 203–215.

Dobin A, Davis CA, Schlesinger F *et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.

Dorus S, Evans PD, Wyckoff GJ, Choi SS, Lahn BT (2004) Rate of molecular evolution of the seminal protein gene SEMG2 correlates with levels of female promiscuity. *Nature Genetics*, **36**, 1326–1329.

Eden E, Lipson D, Yogev S, Yakhini Z (2007) Discovering motifs in ranked lists of DNA sequences. *PLoS Computational Biology*, **3**, 508–522.

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

Ellegren H (2013) The evolutionary genomics of birds. *Annual Review of Ecology, Evolution, and Systematics*, **44**, 239–259.

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

Ellegren H, Smeds L, Burri R *et al.* (2012) The genomic landscape of species divergence in Ficedula flycatchers. *Nature*, **491**, 756–760.

Flicek P, Ahmed I, Amode MR *et al.* (2013) Ensembl 2013. *Nucleic Acids Research*, **41**, D48–D55.

Grabherr MG, Haas BJ, Yassour M *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, **29**, 644–652.

Grath S, Parsch J (2012) Rate of amino acid substitution is influenced by the degree and conservation of male-biased transcription over 50 Myr of *Drosophila* evolution. *Genome Biology and Evolution*, **4**, 346–359.

Harrison PW, Jordan GE, Montgomery SH (2014) swamp: sliding window alignment masker for paml. *Evolutionary Bioinformatics Online*, **10**, 1–8.

Harrison PW, Wright AE, Zimmer F *et al.* (in press) Sexual selection drives evolution and rapid turnover of male-biased genes. PNAS.

Hartl DL, Clark AG (2007) *Principles of Population Genetics*. Sinauer Associates, Inc., Sunderland, Massachusetts.

Holm S (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**, 65–70.

Hope ACA (1968) A simplified Monte Carlo significance test procedure. *Journal of the Royal Statistical Society Series B, Statistical Methodology*, **30**, 582–598.

Hvilsom C, Qian Y, Bataillon T *et al.* (2012) Extensive X-linked adaptive evolution in central chimpanzees. *Proceedings of the National Academy of Sciences of the USA*, **109**, 2054–2059.

Itoh Y, Replogle K, Kim Y-H, Wade J, Clayton DF, Arnold AP (2010) Sex bias and dosage compensation in the zebra finch versus chicken genomes: general and specialized patterns among birds. *Genome Research*, **20**, 512–518.

Khaitovich P, Hellmann I, Enard W *et al.* (2005) Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science*, **309**, 1850–1854.

Kimura M, Ohta T (1971) On the rate of molecular evolution. *Journal of Molecular Evolution*, **1**, 1–17.

Kirkpatrick M, Hall DW (2004a) Male-biased mutation, sex linkage, and the rate of adaptive evolution. *Evolution*, **58**, 437–440.

Kirkpatrick M, Hall DW (2004b) Sexual selection and sex linkage. *Evolution*, **58**, 683–691.

Knief U, Schielzeth H, Kempenaers B, Ellegren H, Forstmeier W (2012) QTL and quantitative genetic analysis of beak morphology reveals patterns of standing genetic variation in an Estrildid finch. *Molecular Ecology*, **21**, 3704–3717.

Koboldt DC, Chen K, Wylie T *et al.* (2009) varscan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, **25**, 2283–2285.

Koboldt DC, Zhang Q, Larson DE *et al.* (2012) varscan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, **22**, 568–576.

Langley CH, Stevens K, Cardeno C *et al.* (2012) Genomic variation in natural populations of *Drosophila melanogaste*r. *Genetics*, **192**, 533–598.

Laporte V, Charlesworth B (2002) Effective population size and population subdivision in demographically structured populations. *Genetics*, **162**, 501–519.

Li B, Dewey CN (2011) rsem: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.

Li H, Handsaker B, Wysoker A *et al.* (2009) The sequence alignment/map format and samtools. *Bioinformatics*, **25**, 2078–2079.

Lohse M, Bolger AM, Nagel A *et al.* (2012) robina: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Research*, **40**, W622–W627.

Mank JE (2009) The W, X, Y and Z of sex chromosome dosage compensation. *Trends in Genetics*, **25**, 226–233.

Mank JE, Ellegren H (2009) All dosage compensation is local: Gene-by-gene regulation of sex-biased expression on the chicken Z chromosome. *Heredity*, **102**, 312–320.

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

Mank JE, Hultin-Rosenberg L, Axelsson E, Ellegren H (2007b) Rapid evolution of female-biased, but not male-biased, genes expressed in the avian brain. *Molecular Biology and Evolution*, **24**, 2698–2706.

Mank JE, Vicoso B, Berlin S, Charlesworth B (2010a) Effective population size and the Faster-X Effect: empirical results and their interpretation. *Evolution*, **64**, 663–674.

Mank JE, Nam K, Ellegren H (2010b) Faster-Z evolution is predominantly due to genetic drift. *Molecular Biology and Evolution*, **27**, 661–670.

Mank JE, Nam K, Brunström B, Ellegren H (2010c) Ontogenetic complexity of sexual dimorphism and sex-specific selection. *Molecular Biology and Evolution*, **27**, 1570–1578.

McCarthy DJ, Chen Y, Smyth GK (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, **40**, 4288–4297.

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the ADH locus in *Drosophila*. *Nature*, **351**, 652–654.

Meisel RP (2011) Towards a more nuanced understanding of the relationship between sex-biased gene expression and rates of protein coding sequence evolution. *Molecular Biology and Evolution*, **28**, 1893–1900.

Meisel RP, Connallon T (2013) The faster-X effect: integrating theory and data. *Trends in Genetics*, **29**, 537–544.

Moller AP (1988) Testes size, ejaculate quality and sperm competition in birds. *Biological Journal of the Linnean Society*, **33**, 273–283.

Moller AP (1991) Sperm competition, sperm depletion, paternal care, and relative testis size in birds. *The American Naturalist*, **137**, 882–906.

Moller AP, Briskie JV (1995) Extra-pair paternity, sperm competition and the evolution of testis size in birds. *Behavioral Ecology and Sociobiology*, **36**, 357–365.

Montgomery SH, Capellini I, Venditti C, Barton RA, Mundy NI (2011) Adaptive evolution of four microcephaly genes and the evolution of brain size in anthropoid primates. *Molecular Biology and Evolution*, **28**, 625–638.

Pagel M (1999) Inferring the historical patterns of biological evolution. *Nature*, **401**, 877–884.

Pagel M, Meade A, Barker D (2004) Bayesian estimation of ancestral character states on phylogenies. *Systematic Biology*, **53**, 673–684.

Parsch J, Ellegren H (2013) The evolutionary causes and consequences of sex-biased gene expression. *Nature Reviews. Genetics*, **14**, 83–87.

Patefield WM (1981) ALGORITHM AS159. An efficient method of generating r x c tables with given row and column totals. *Applied Statistics*, **30**, 91–97.

Perry JC, Harrison PW, Mank JE (2014) The ontogeny and evolution of sex-biased gene expression in *Drosophila melanogaster*. *Molecular Biology and Evolution*, **31**, 1206–1219.

Petrie M, Krupa A, Burke T (1999) Peacocks lek with relatives even in the absence of social and environmental cues. *Nature*, **401**, 155–157.

Pitcher TE, Dunn PO, Whittingham LA (2005) Sperm competition and the evolution of testes size in birds. *Journal of Evolutionary Biology*, **18**, 557–567.

Pointer MA, Harrison PW, Wright AE, Mank JE (2013) Masculinization of gene expression is associated with exaggeration of male sexual dimorphism. *PLoS Genetics*, **9**, 1–9.

Pool JE, Nielsen R (2007) Population size changes reshape genomic patterns of diversity. *Evolution*, **63**, 3001–3006.

Quinn EM, Cormican P, Kenny EM *et al.* (2013) Development of strategies for SNP detection in RNA-Seq data: Application to lymphoblastoid cell lines and evaluation using 1000 genomes data. *PLoS One*, **8**, e58815.

R Core Team (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Ramm SA, Oliver PL, Ponting CP, Stockley P, Emes RD (2008) Sexual selection and the adaptive evolution of Mammalian ejaculate proteins. *Molecular Biology and Evolution*, **25**, 207–219.

Reeve HK, Pfennig DW (2003) Genetic biases for showy males: are some genetic systems especially conducive to sexual selection? *Proceedings of the National Academy of Sciences of the USA*, **100**, 1089–1094.

Rice WR (1984) Sex chromosomes and the evolution of sexual dimorphism. *Evolution*, **38**, 735–742.

Robinson MD, Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, **11**, R25.

Robinson MD, McCarthy DJ, Smyth GK (2010) EDGER: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Ruokonen M, Kvist L, Lumme J (2000) Close relatedness between mitochondrial DNA from seven Anser goose species. *Journal of Evolutionary Biology*, **13**, 532–540.

Sackton TB, Corbett-Detig RB, Nagaraju J, Vaishna L, Arunkumar KP, Hartl DL (2014) Positive selection drives Faster-Z evolution in silkmoths. *Evolution*, **68**, 2331–2342.

Saether SA, Saetre G-P, Borge T *et al.* (2007) Sex chromosome-linked species recognition and evolution of reproductive isolation in flycatchers. *Science*, **318**, 95–97.

Saetre GP, Borge T, Lindroos K *et al.* (2003) Sex chromosome evolution and speciation in Ficedula flycatchers. *Proceedings of the Royal Society B-Biological Sciences*, **270**, 53–59.

Schielzeth H, Kempenaers B, Ellegren H, Forstmeier W (2012) QTL linkage mapping of zebra finch beak color shows an oligogenic control of a sexually selected trait. *Evolution*, **66**, 18–30.

Singh N, Larracuente A, Clark A (2008) Constrasting the efficacy of selection on the X and autosomes in *Drosophila*. *Molecular Biology and Evolution*, **25**, 454–467.

Skinner BM, Robertson LBW, Tempest HG *et al.* (2009) Comparative genomics in chicken and Pekin duck using FISH mapping and microarray analysis. *BMC Genomics*, **10**, 357.

Stiglec R, Ezaz T, Graves JAM (2007) A new look at the evolution of avian sex chromosomes. *Cytogenetic and Genome Research*, **117**, 103–109.

Thornton K, Long M (2005) Excess of amino acid substitutions relative to polymorphism between X-linked duplications in *Drosophila melanogaster*. *Molecular Biology and Evolution*, **22**, 273–284.

Toups MA, Pease JB, Hahn MW (2011) No excess gene movement is detected off the avian or lepidopteran Z chromosome. *Genome Biology and Evolution*, **3**, 1381–1390.

van Tuinen M, Dyke GJ (2004) Calibration of galliform molecular clocks using multiple fossils and genetic partitions. *Molecular Phylogenetics and Evolution*, **30**, 74–86.

van Tuinen M, Hedges SB (2001) Calibration of avian molecular clocks. *Molecular Biology and Evolution*, **18**, 206–213.

Uebbing S, Künstner A, Mäkinen H, Ellegren H (2013) Transcriptome sequencing reveals the character of incomplete dosage compensation across multiple tissues in flycatchers. *Genome Biology and Evolution*, **5**, 1555–1566.

Vicoso B, Charlesworth B (2009) Effective population size and the Faster-X effect: an extended model. *Evolution*, **63**, 2413–2426.

Vicoso B, Emerson JJ, Zektser Y, Mahajan S, Bachtrog D (2013a) Comparative sex chromosome genomics in snakes: differentiation, evolutionary strata, and lack of global dosage compensation. *PLoS Biology*, **11**, e1001643.

Vicoso B, Kaiser VB, Bachtrog D (2013b) Sex-biased gene expression at homomorphic sex chromosomes in emus and its implication for sex chromosome evolution. *Proceedings of the National Academy of Sciences of the USA*, **110**, 6453–6458.

Wang B, Ekblom R, Bunikis I, Siitari H, Hoglund J (2014a) Whole genome sequencing of the black grouse (*Tetrao tetrix*): reference guided assembly suggests faster-Z and MHC evolution. *BMC Genomics*, **15**, 180.

Wang Z, Zhang J, Yang W *et al.* (2014b) Temporal genomic evolution of bird sex chromosomes. *BMC Evolutionary Biology*, **14**, 250.

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, **7**, 256–276.

Wright AE, Mank JE (2013) The scope and strength of sex-specific selection in genome evolution. *Journal of Evolutionary Biology*, **26**, 1841–1853.

Xu K, Oh S, Park T, Presgraves DC, Yi SV (2012) Lineage-specific variation in slow- and fast-X evolution in primates. *Evolution*, **66**, 1751–1761.

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, **24**, 1586–1591.

A.E.W., M.A.P. and J.E.M. were involved in data collection. A.E.W. and J.E.M. designed the research. A.E.W., P.W.H., F.Z., S.H.M. and J.E.M. analysed the data and wrote the paper.

Data accessibility

Illumina raw reads: SRA: PRJNA271731.

Trinity assembly, RPKM data, sequence alignments and SNP data: Dryad: doi:10.5061/dryad.4gv50.

Supporting information

Additional supporting information may be found in the online version of this article.

Table S1 Assembly statistics for each Galloanserae species.

Table S2 Assembly statistics for each sample.

Table S3 Number of nonsynonymous and synonymous polymorphic and fixed sites at a minor allele frequency threshold of 0.15.

Table S4 Number of nonsynonymous and synonymous polymorphic and fixed sites at a minor allele frequency threshold of 0.25.

Table S5 Phylogenetically controlled regression analyses between Faster-Z and log sperm number, and residual testes weight.

Table S6 Nucleotide diversity estimates of the Z chromosome and autosomes.

Table S7. Effective population size estimates of the Z chromosome and autosomes.

Table S8 Effective population size estimates of the Z chromosome and autosomes calculated using Watterson's estimation of theta.

Table S9 Phylogenetically controlled regression analyses between different measures of NEZ/NEA and residual testes weight, log sperm number, and Faster-Z.

Table S10 Site model test results for contigs with signatures of positive selection.

Table S11 pN, Ps and pN/pS for Z-linked and autosomal genes across Galloanserae species with a minor allele frequency of 0.25.

Table S12 McDonald Kreitman test results.

Table S13 Significant differences between nonsynonymous and synonymous polymorphism on the Z chromosome and autosomes with a minor allele frequency threshold of 0.25.

Table S14 Faster-Z Effect across expression classes.

Table S15 Differences in Faster-Z Effect between expression classes.

The following article was first published in

Nature Communications

Wright AE, Dean R, **Zimmer F**, Mank JE (2016) How to make a sex chromosome. *Nature Communications* 7:12087
doi: 10.1038/ncomms12087

nature COMMUNICATIONS

REVIEW

Received 11 Jan 2016 | Accepted 27 May 2016 | Published 4 Jul 2016

DOI: 10.1038/ncomms12087

OPEN

How to make a sex chromosome

Alison E. Wright[1], Rebecca Dean[1], Fabian Zimmer[1] & Judith E. Mank[1]

Sex chromosomes can evolve once recombination is halted between a homologous pair of chromosomes. Owing to detailed studies using key model systems, we have a nuanced understanding and a rich review literature of what happens to sex chromosomes once recombination is arrested. However, three broad questions remain unanswered. First, why do sex chromosomes stop recombining in the first place? Second, how is recombination halted? Finally, why does the spread of recombination suppression, and therefore the rate of sex chromosome divergence, vary so substantially across clades? In this review, we consider each of these three questions in turn to address fundamental questions in the field, summarize our current understanding, and highlight important areas for future work.

Sex chromosomes have evolved independently many times throughout the eukaryotes, and represent a remarkable case of genomic convergence, as unrelated sex chromosomes share many properties across distant taxa[1–3]. Sex chromosomes evolve after recombination is halted between a homologous pair of chromosomes[4,5], leading to a cascade of non-adaptive and adaptive processes that produce distinct differences between the X and Y (or Z and W) chromosomes.

Owing to detailed studies in *Drosophila*[6–8] and mammals[9–11], we have a nuanced understanding of the consequences of arrested recombination[1,4,7,8]. The non-recombining Y and W chromosomes become highly heterochromatic (see Box 1 for a glossary) and experience profound levels of gene loss even as the X and Z chromosomes remain functional[1,12–14]. Sex chromosomes have been the focus of intense study and are an important model for understanding the consequences of recombination suppression[12,15]. It is clear that the loss of recombination triggers a host of evolutionary processes, including Muller's Ratchet, background selection and genetic hitchhiking, reviewed in ref. 16, that lead to the loss of gene activity and pseudogenization (detailed in Box 2). This work makes very clear the evolutionary consequences of halting recombination between the sex chromosomes.

Why recombination is suppressed in the first place is less clear, as the chromosomes that determine sex in many organisms with genetic sex determination never progress to heteromorphic sex chromosomes. For example, a single missense single nucleotide polymorphism in the coding region of the *Amhr2* locus appears to control sex in the tiger pufferfish (*Takifugu rupripes*)[17], but recombination is not restricted around this sex-determining gene and there is no evidence of divergence beyond this single nucleotide between the proto-X or proto-Y. Similarly, despite considerable age, the sex chromosomes in many clades (including ratite birds[18,19], pythons[20] and European tree frogs[21]) have failed to develop substantial heteromorphism, and remain largely identical.

These observations indicate that recombination suppression and sex chromosome divergence are not inevitable consequences of genetic sex determination, leading to three questions at

[1] Department of Genetics, Evolution and Environment University College London, London WC1E 6BT UK. Correspondence should be addressed to J.E.M. (email: Judith.Mank@ucl.ac.uk).

Box 1 | Glossary.

Achiasmate: Complete suppression of recombination in one sex, typically the heterogametic sex. Observed in *Drosophila* and Lepidoptera, among others.

Dioecy: Botanical term for separate male and female flowers in different individuals. Similar to gonochorism in animals.

Dosage compensation: Gene regulation mechanism on the sex chromosomes to correct for differences in gene dose for the X or Z chromosome between the homogametic and heterogametic sexes (Fig. 2). A consequence of dosage compensation is that gene dose is equalized between males and females.

Female heterogamety: Sex chromosome type where females have a ZW karyotype, and males a ZZ karyotype. Present in birds, lepidoptera, snakes and anguillid eels.

Gonochorism: Animal term for separate sexes. Similar to dioecy in plants.

Gynodioecy: Male sterile individuals and hermaphrodites in the same population.

Heterochiasmy: Sex-specific variation in recombination rates.

Heteromorphic sex chromosomes: Sex chromosomes that are karyotypically highly distinct from one another. In these cases, the X and Y (or Z and W) chromosomes show major differences in size and gene content.

Homomorphic sex chromosomes: Where the X and Y (or Z and W) chromosomes exhibit few differences from each other in size and gene content, and are difficult or impossible to distinguish from karyotype data alone.

Hermaphrodite: Male and female reproductive organs in the same individual.

Male heterogamety: Type of sex chromosome system where females karyotype is XX, and male karyotype is XY. Observed in mammals, *Drosophila*, salmon as well as many beetles.

Pseudo-autosomal region: Regions where recombination persists between the X and Y (or Z and W) sex chromosomes. These regions, identical in both sexes, aid chromosome pairing during meiosis and ensure proper segregation.

Pseudogene: DNA sequences that once encoded protein sequences, but which are no longer transcribed into messenger RNA (mRNA) in a way that translates to functional protein.

Stratum: Region on the sex chromosomes where recombination has been suppressed. Strata can be identified by spatial clusters of X-Y or Z-W orthologs with similar divergence estimates.

the heart of sex chromosomes evolution. First, why do sex chromosomes stop recombining? Second, how is recombination suppression achieved? Third, why does the spread of recombination suppression, and therefore the rate of sex chromosome divergence, vary so substantially across clades?

The implications of these questions go far beyond sex chromosome research *per se*. Recombination rate has long been known to be a critical factor in the ability of a genomic region to respond to selection. Dobzhansky and colleagues[22–25] noted that halting recombination can permanently link co-adapted gene complexes (recently renamed supergenes) within populations. These supergenes are then transmitted as a unit, allowing for complex adaptions spanning multiple loci. More recently, the importance of recombination has resurfaced in evolutionary biology with several key examples in a range of species implicating recombination suppression as a crucial component of complex phenotypic adaptation[26–29] and speciation[30]. The study of sex chromosomes therefore offers a route to understand the interplay between recombination, selective forces and adaptation, with broad implications across multiple fields of evolutionary genetics.

Why do sex chromosomes stop recombining

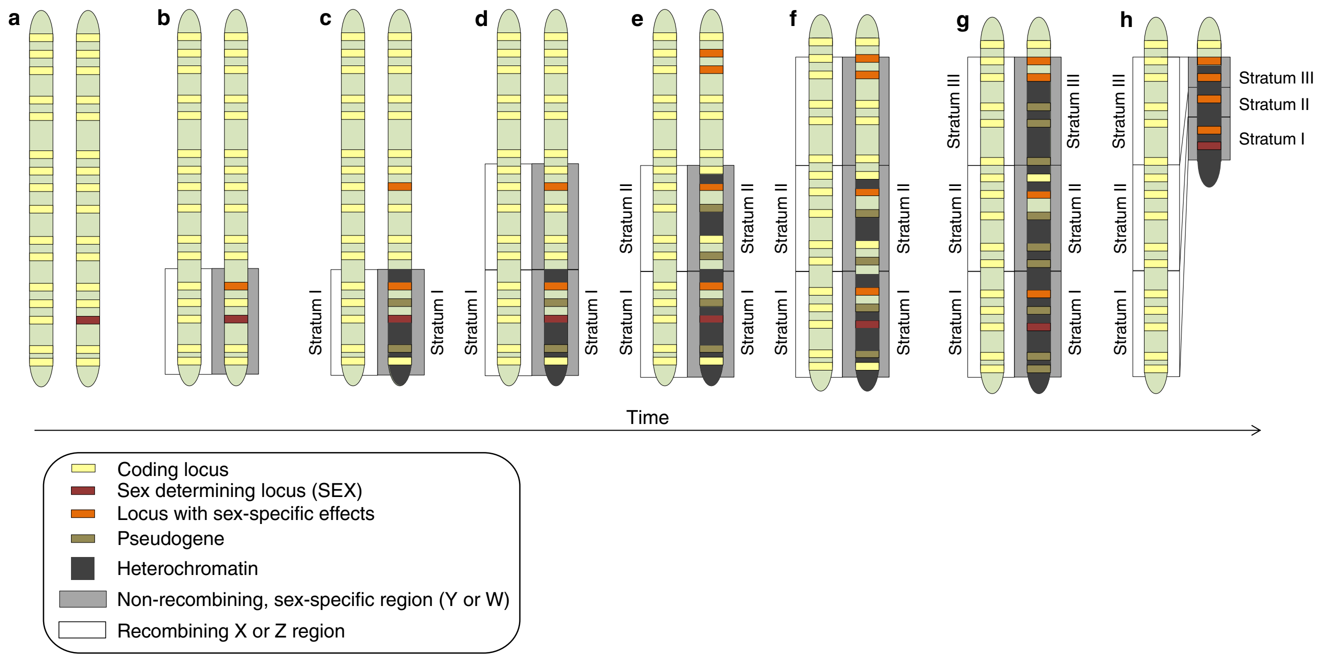
The sexual conflict model of sex chromosome evolution. The most commonly accepted theory of sex chromosome evolution[14,31,32] predicts that recombination will be selected against in the region between a sex-determining gene and a nearby gene with sex-specific effects (Box 2). This theory was based in part on early studies of colouration genetics in the guppy, *Poecilia reticulata*[33], which demonstrated that many genes underlying male colouration are Y-linked. Colouration genes are sexually antagonistic—they benefit males through increased reproductive success but are detrimental to both sexes due to increased predation. For males, the benefits of increased mating opportunities outweigh the costs when predation pressures are not too high. In contrast, females gain no benefit from displaying bright colours to offset increased predation, as males are not attracted to ornamented females. Linkage between the allele that confers maleness at the sex determining locus and the allele for bright coloration at a nearby locus creates a male supergene—the allele determining maleness is always co-inherited with the linked allele, which confers a fitness benefit in males. The linkage of these alleles also resolves sexual conflict over colour between males and females, as the colouration allele would no longer be present, and therefore selected against, in females.

Although the sexual conflict model of sex chromosome evolution remains widely accepted, the evidence for or against it is remarkably slim. Non-adaptive alternatives have been suggested as well[34,35], but also lack definitive evidence. Clear empirical evidence to support the sexual conflict theory of sex chromosome evolution is limited in part because the main model species for empirical studies of sex chromosome evolution exhibit highly derived X and Y chromosomes, requiring substantial extrapolation to infer the initial stages of divergence.

Importantly, it can be difficult in ancient systems to differentiate cause from consequence. For example, the gene content of the Y chromosome has been interpreted as supporting the role of sexual conflict in sex chromosome evolution. The Y chromosome in mammals[36] and *Drosophila*[37,38], as well as the analogous W chromosome in birds[39], contains loci essential to sex-specific fitness, which might have been sexually antagonistic before they became sex-limited (linked to the Y or W chromosome). However, although sexual conflict over these loci could have catalyzed sex chromosome divergence through selection for recombination suppression (supporting the sexual conflict model), these genes could just as easily have relocated after recombination halted[40]. In support of this latter explanation, there is evidence of strong selection for the relocation of male-benefit gene duplicates to the Y chromosome in *Drosophila*[40]. Alternatively, these genes may have developed sex-specific functions after the sex chromosomes diverged, as there is also evidence that loci on sex chromosomes adapt to their sex-specific environment once recombination ceases[41]. Y-linked loci would therefore be more likely to adopt male-specific functions after recombination with the X chromosome is halted, but these functions would not drive recombination suppression itself.

Evidence from sex chromosome systems at earlier stages of divergence is therefore key to understanding why sex chromosomes evolve, and there are a wealth of systems with early stage sex chromosomes including *Anolis* lizards[42,43], anurans[21,44,45], snakes[46], fish[47], many plants[48–51], among numerous others[2]. However, although these systems have revealed several important characteristics of early stage sex chromosome evolution, the difficulty in identifying sexually antagonistic alleles at the molecular level has hampered direct empirical tests of the sexual conflict model. Indirect evidence for the sexual conflict model comes from the three-spine stickleback (*Gasterosteus aculeatus*), where a neo-sex chromosome fusion in the Sea of Japan population may have been driven, at least in part, by sexual conflict[52]. However, recombination suppression has not spread across the added region, suggesting that linkage between the

Box 2 | Theoretical model of sex chromosome differentiation.


a
b
c
d
e
f
g
h
Stratum I
Stratum II
Stratum III
Time
Coding locus
Sex determining locus (SEX)
Locus with sex-specific effects
Pseudogene
Heterochromatin
Non-recombining, sex-specific region (Y or W)
Recombining X or Z region

Sex chromosomes evolve from autosomes, initially with the acquisistion of a sex determining locus (**a**). Emergence of sexually antagonistic alleles at loci in close proximity to the sex determining locus selects for recombination suppression between the X and Y or Z and W chromosome (**b**), resulting in Stratum I, which is increasingly heterochromatinized. Once recombination is halted on the Y or W chromosome genes without sex-specific benefits are often pseudogenized. The non-recombining region can expand with the acquisition of additional sexually antagonistic alleles and further recombination suppression, leading to additional strata—spatial clusters of X-Y or Z-W orthologs with similar divergence estimates, observed in mammals[9], birds[39,109], fish [67,94] and plants[48,66], which also undergo loss of gene function and heterochromatinization (**d**–**g**). The lack of recombination leads to accumulation of repetitive DNA, which can lead to a short-term increase in the size of the Y or W, but which typically results in large-scale deletions, a large reduction in physical size of the sex-limited chromosome, and highly heteromorphic sex chromosomes (**h**)[7,65].

Sex chromosomes may emerge in a somewhat different way in species where one sex or the other lack recombination at all. Referred to as achiasmy, this occurs in a range of species, most notably *Drosophila*[110] and Lepidoptera[111,112], but also within Hemiptera[113], Heteroptera[114–116] and Orthoptera[117] and with restricted distributions in several other taxa[61]. In these cases, if achiasmy precedes the emergence of a nascent sex determining locus, linkage between two or more loci is not required for recombination to cease between the emergent sex chromosomes. The advent of a sex determining allele automatically makes the entire chromosome sex-limited and therefore non-recombining. In these cases, there are no discernible strata.

sexually antagonistic locus and the sex determining locus may not explain the fusion event[53]. Similarly, a sexually antagonistic colouration pattern has been mapped to the W chromosome in some cichlids[54]; however, given the dynamic and polygenic nature of sex determination in cichlids[55], it is not clear whether W-linkage predates sex chromosome evolution or that linkage of the coloration locus to the sex determining gene led to recombination suppression.

Transitions from hermaphroditism to sex chromosomes. The theory of sex chromosome evolution articulated above assumes that the separation of the sexes, called gonochorism in animals and dioecy in plants, predates the evolution of sex chromosomes. Because of this assumption, the theory is in many ways more applicable to animals, which are more often gonochoristic. Dioecy is rare in plants, which restricts the evolution of sex chromosomes to fewer taxa. In flowering plants (angiosperms), only 5–6% of all species have separate male and female genders[56]. Of the dioecious angiosperms, only a small number have been shown to possess sex chromosomes of which roughly half are homomorphic[56,57]. However, without detailed genetic analysis, homomorphic sex chromosomes are difficult to identify. As a result, there may be many cryptic homomorphic species where the sex chromosomes are karyotypically indistinguishable and just waiting to be discovered.

In plants and other systems where sex chromosomes are associated with transitions from hermaphroditism to separate sexes, sex chromosome formation may take a slightly different route than in species with ancestral separate sexes. In this case, the dominant model[58] predicts that separate male- and female-sterile mutations on the same chromosome cause the shift from hermaphroditism to dioecy through an intermediate phase of gynodioecy. Once these mutations have occurred and reached sufficient frequency in the population, recombination suppression between them prevents reversal back to hermaphroditism, leading to the evolution of sex chromosomes. Recent evidence from wild strawberry[59] and papaya[49,60] has provided insight into these early stages of sex chromosome evolution in plants and the availability of genomic tools will help us understand how recombination is suppressed between feminizing and masculinizing alleles.

How is recombination halted between the sex chromosomes

Regardless of why sex chromosomes originate, the process of sex chromosome evolution necessitates halting recombination between the nascent X and Y in males, or Z and W in females. Therefore, sex chromosome evolution at the most basic level requires sex-specific recombination patterns on the sex chromosomes. Recombination varies substantially in males and females, both in frequency and in specific hotspots, referred to as

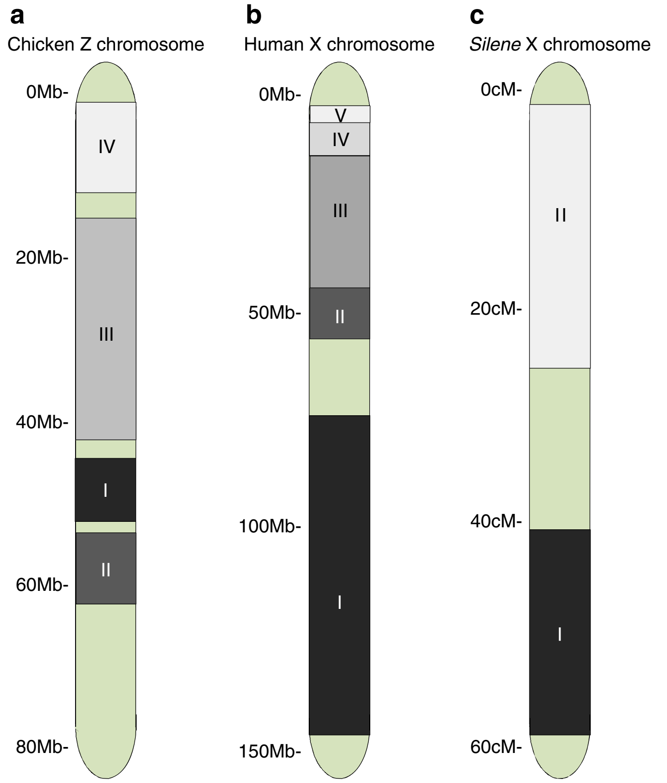


Figure 1 | Sex chromosome strata. Many plants and animals show evidence of strata, spatial clusters of X-Y, or Z-W, orthologs with similar divergence estimates. These spatial clusters are consistent with inversion events instantaneously halting recombination for all the encompassed loci. As inversions are proposed to occur in a stepwise process, strata differ in the length of time over which recombination has been suppressed. Therefore, orthologs with the largest neutral sequence divergence reside in the oldest stratum (shown in black), whereas those with the greatest sequence similarity are located in the youngest stratum (shown in white). The chicken Z chromosome (**a**) is comprised of at least four strata, formed over 130 million years[68] and the human X chromosome (**b**) is comprised of at least five strata[105], although some recent analyses support six or more strata[106,107]. The *Silene* X and Y chromosomes (**c**) diverged more recently and there is evidence for two strata over 10 million years[66]. However, it is possible that orthology-based approaches underestimate the number of strata (regions unassigned to strata shown in green). For example, in highly degenerated regions, often all of the Y or W loci have decayed and no orthologs remain. In these cases, alternative methods have been used to identify additional strata[92,108].

heterochiasmy. An extreme example of this is achiasmy, where recombination only occurs in one sex[61].

Achiasmy may either precede or follow emergence of a nascent sex determining locus[62,63], and in either case, can accelerate sex chromosome divergence. For example, in an achiasmate species, the emergence of a nascent sex determining factor leads to instantaneous recombination suppression along the entire length of the sex chromosomes. Similarly, when achiasmy follows quickly after the emergence of a nascent sex determining factor, recombination suppression also occurs along the entire length of the sex chromosomes. Only when achiasmy evolves in systems with highly differentiated sex chromosomes would it not be expected to foster sex chromosome divergence. As a result, the sex chromosomes of achiasmate species tend to have a single heteromorphic stratum, as the emergence of a new sex determining allele causes the entire sex chromosome to start to diverge[64].

Box 3 | Sex-specific recombination.

Recombination rates show substantial variation within the genome[73,77,118,119], within species[71,74,80] as well as across related species[73,78,79]. Importantly for sex chromosome evolution, there are often also differences between males and females in recombination rate, and sex differences in recombination rates are thought to occur in >75% of recombining species. In many cases, the magnitude of the difference can be very large[62,63,120]. In general, males tend to have lower rates of recombination than females during meiosis and this pattern is independent of male or female heterogamety[63].

Sex-specific recombination rates, and in particular local sex-specific recombination cold-spots, may be important for initiating sex chromosome degeneration. Furthermore, sexual dimorphism in recombination could promote the spread of sexually antagonistic alleles, as low recombination in the sex that benefits from the sexually antagonistic genes keeps favourable sexually antagonistic combinations together[121], which in turn could drive expansion of the non-recombining region and progressive sex chromosome evolution. Yet the evolutionary forces and molecular mechanisms driving sex-specific recombination are relatively unknown. Possible selective forces causes include stronger haploid selection in males than females[63] and various forms of epistatic selection[62]. Understanding the mechanisms underlying recombination cessation, what causes inter- and intra-specific recombination rates, and whether achiasmate recombination is a cause or consequence of sex chromosome evolution will provide greater understanding of sex chromosome evolution.

In species where both sexes recombine, some mechanism is needed to block recombination between the sex determining gene and nearby genes with sex-specific effects in the heterogametic sex. Chromosomal inversions spanning the sex determining locus and nearby sexually antagonistic loci are often assumed to halt recombination and therefore to drive sex chromosome divergence[65]. There is circumstantial evidence implicating inversions in sex chromosome evolution. For example, sex chromosomes in many animals and plants show evidence of strata, spatial clusters of X-Y or Z-W orthologs with similar divergence estimates (Fig. 1)[10,20,48,66–68]. These spatial clusters are consistent with inversion events instantaneously halting recombination for all the encompassed loci. However, reports from nascent sex chromosomes suggest that recombination suppression is initially heterogeneous across the sex chromosomes[53,69,70], implying that recombination suppression evolves initially by another, uneven mechanism, inconsistent with large-scale inversions.

Recombination is dynamic and heterogeneous, and the rate of recombination varies extensively throughout the genome and between the sexes[63,71]. For species where both sexes recombine, local sex-specific recombination rates may be important initially in sex chromosome divergence, although the mechanism for sex-specific heterochiasmy is not yet known (Box 3). Importantly, regardless of the mechanism, once recombination has been halted in the heterogametic sex, selection to maintain gene order is abolished[72] and inversions are less likely to be selected against. Relaxed selection against inversions suggests that inversions might follow recombination suppression. Therefore, it remains unclear whether inversions catalyze or are a consequence of halting recombination between sex chromosomes.

Recent work on recombination evolution has suggested that sequence characteristics, namely binding motifs and structural traits, can exhibit short-term evolutionary dynamics that can lead to rapid shifts in local recombination rates[73–75]. Although not present in all species[76,77], when they are associated with recombination, rapid changes in these motifs lead to differences in recombination rates in specific genomic locations among closely related species[73,78,79], and even among conspecific populations[71,74,80]. The role of structural modifications and binding motifs in sex chromosome evolution, as well as other genetic and epigenetic mechanisms (detailed in ref. 81), have yet to be explored, but these mechanisms offer plausible alternatives to inversions in driving recombination suppression.

Why do sex chromosomes diverge at such different rates

Homomorphic sex chromosomes are curiously common. Many organisms with genetic sex determination lack heteromorphic sex chromosomes, indicating that the non-recombining region has not spread significantly beyond the sex determining locus. Examples of animal systems with homomorphic sex chromosomes include the pufferfish[17], ratite birds[18,19], pythons[20] and European tree frogs[21]. Also, many dioecious species of flowering plants possess homomorphic sex chromosomes[82]. The reasons why sex chromosomes might remain largely undifferentiated are not well understood, but here we suggest five possible explanations.

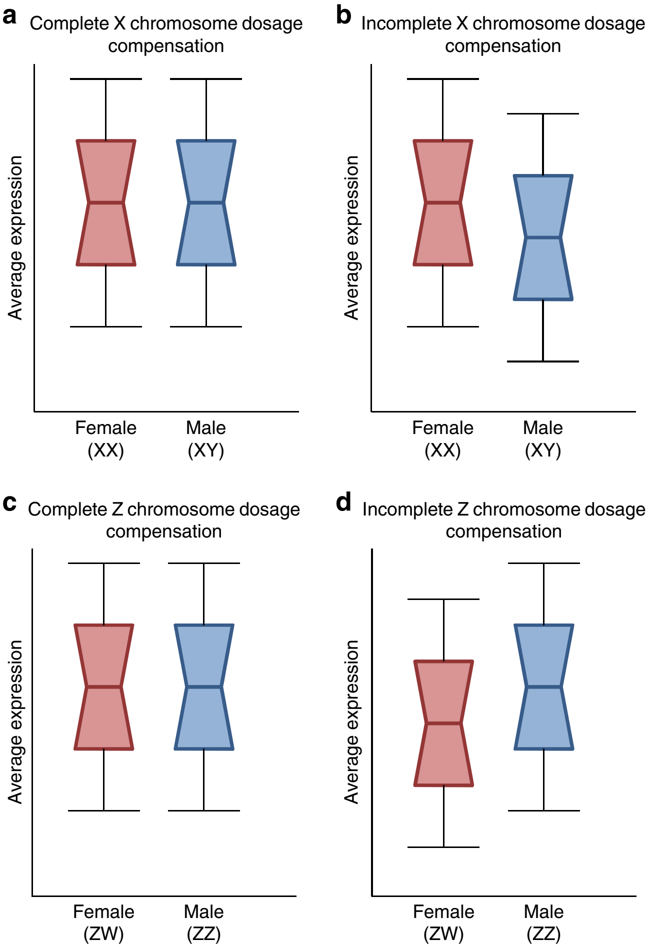


Figure 2 | Cartoon illustration of sex chromosome dosage compensation. The decay of Y and W chromosome gene content leads to differences in gene dose (the number of gene copies) between the sexes. In male heterogamety (**a**,**b**) males have one half of the dose of all X-linked genes lost from the Y chromosome. In some cases, this difference in gene dose has led to the evolution of complete sex chromosome dosage compensation (**a**), where a mechanism acts across the chromosome to balance out the differences in gene dose, and as a consequence, the average expression for X-linked genes is equal in males and females. In many other cases (**b**), only some genes on the X are compensated, and the average expression from the X chromosome is less in males than females. In female heterogamety (**c**,**d**) females have one half of the dose of all Z-linked genes lost from the W chromosome. In some cases, this difference in gene dose has led to the evolution of complete sex chromosome dosage compensation (**c**), but in many other cases (**d**), only some genes on the Z are compensated, and the average expression from the Z chromosome is less in females than males.

Age. First, some homomorphic sex chromosomes are young and may be in the early stages of degeneration, for example in papaya[49,60]. However, in many species, the sex chromosomes are old and yet have not degenerated, such as in European tree frogs[21], pythons[20] and ratite birds[19]. Thus, we must conclude that age is not always an accurate predictor of the relative size of the non-recombining region, and therefore of overall sex chromosome divergence.

Relative length of haploid phase. Some organisms have a long haploid phase, resulting in strong haploid purifying selection acting to maintain gene activity on the Y chromosome[70,83,84]. In species where haploid selection is more limited, many genes on the Y or W chromosome are sheltered in the diploid phase by the copy on the X or Z chromosome, and purifying selection may only act on dosage sensitive genes to maintain sufficient gene activity. Therefore, we might expect slower W or Y degeneration in species where haploid selection is more pervasive, such as algae and plants, compared with species where it is less widespread, such as animals. Similarly, some animals have a much reduced haploid phase in females compared to males, and this might retard W chromosome degeneration compared to that of Y chromosomes[63].

Sex chromosome dosage compensation. After recombination has been halted between the sex chromosomes, the non-recombining Y or W chromosome decays[85]. A consequence of this degeneration is that gene dose is reduced on the X and Z chromosomes relative to the autosomes in the heterogametic sex. This imbalance in gene expression is often thought to be detrimental, and upsets the biochemical stoichiometry of interacting gene products. These deleterious effects were hypothesized to drive the evolution of dosage compensation mechanisms in order to restore ancestral diploid expression levels[86]. The extent of dosage compensation varies significantly across taxa[87], and although some species exhibit complete sex chromosome dosage compensation, many more show incomplete compensation (reviewed in refs 87,88, shown in Fig. 2). The factors underlying this variation are not at all clear and may include sexual conflict over optimal gene expression[89], as well as variation in effective population size and male-biased mutation rates.

Much of our understanding of Y chromosome decay comes from the neo-sex chromosomes in *Drosophila* and the X-added region of the eutherians. In both these cases, an existing system of complete dosage compensation quickly spread onto the expanded X chromosome[90,91]. The spread of an existing mechanism of dosage compensation onto a neo-sex chromosome would reduce the power of purifying selection to maintain gene activity on dosage sensitive neo-Y orthologs, in turn leading to an acceleration of neo-Y chromosome decay.

The slow rate of gene decay recently observed on the W chromosome in birds[92] provides a stark contrast to the *Drosophila* and eutherian Y, and it was recently suggested that this difference is largely due to the opposing effects of male-biased mutation on Y and W chromosomes[1,93]. However, birds have only incomplete sex chromosome dosage compensation[87], raising questions about the generality of the lessons from the *Drosophila* neo-sex chromosomes and the eutherian X-added region, as well as suggesting that the dichotomy between *Drosophila* and eutherians versus birds might not be heterogamety (XY versus ZW), but rather complete versus incomplete dosage compensation. Recent work in sticklebacks, a male heterogametic system with incomplete dosage compensation, indicates that purifying selection remains strong on dosage sensitive Y genes[94]. Therefore it may be that in systems with incomplete dosage compensation, Y or W degeneration might be retarded through purifying selection acting on dosage sensitive genes, and that dosage compensation status may be a major factor underlying differences in sex chromosome degeneration rates.

Sex reversal. Sex reversal, discordance between an individual's phenotypic and genotypic sex, may be important in recombination suppression and sex chromosome evolution. In many ectotherm vertebrates, such as amphibians[95,96] and teleost fish[97], sex reversal results in reproductively viable individuals. Interestingly, because recombination patterns typically follow phenotypic but not genotypic sex, recombination can occur along the full length of the sex chromosomes in individuals with phenotypes that do not match their sex chromosome complement. Even when at very low frequency in the population, sex reversal can prevent sex chromosome divergence and lead to very old homomorphic sex chromosomes[98], as has been shown in frogs[21,99,100].

Sexual conflict. Sexually antagonistic alleles are central to the sexual conflict model of sex chromosome evolution[32], and systems with more sexual conflict experience more rapid expansion of the non-recombining region simply because more loci within the genome, and by extension proximate to the sex determining locus, carry sexually antagonistic alleles[101]. Heteromorphic sex chromosomes might be therefore expected to occur more often in lineages with high levels of sexual conflict and/or sexual dimorphism. However, sexual conflict might also trigger turnover of sex chromosomes[102,103], thereby restarting the process of sex chromosome divergence. It is therefore unclear whether we should expect a direct relationship between the degree of sexual conflict and the size of the non-recombining region.

Conclusion

Three major questions regarding the evolution of sex chromosomes remain unanswered. To answer them, it will be important to move well beyond the main model systems, and develop new study systems at earlier stages of sex chromosome divergence.

Does sexual conflict drive sex chromosome evolution? The role of sexual conflict in driving sex chromosome evolution, although widely accepted, remains fundamentally unknown, largely due to difficulties in identifying sexually antagonistic alleles directly. In order to answer this question, it is important that we develop new study systems with far younger sex chromosomes. Crucially, these study systems will also need to have some phenotypic trait or traits that are known to be sexually antagonistic, with known underlying genetic architecture. Alternatively, experimental evolution of sexual conflict may prove useful in studying changes in sex-specific recombination rates.

How is recombination suppressed between the sex chromosomes? The mechanisms underlying recombination suppression are still largely unknown. Inversions are often assumed to facilitate sex chromosome divergence through recombination suppression, but this assumption is contradicted by the heterogeneity in divergence observed in young sex chromosome systems. Moreover, in old sex chromosome systems, it may be impossible to determine whether inversions catalyze sex chromosome evolution or are a consequence of recombination suppression achieved through other means. This difficulty in differentiating cause and effect again suggests that study systems with nascent sex chromosomes are crucial for understanding the cause of recombination suppression.

Why do rates of sex chromosome divergence vary so significantly across groups? Preliminary evidence suggests that the presence or absence of complete dosage compensation, the relative length of the haploid phase in the life cycle, and the prevalence and fertility of sex reversed individuals might be the largest predictors of the power of purifying selection to maintain gene activity on the sex-limited chromosome, and therefore the rate of gene loss once recombination is halted. The pervasiveness of sexual conflict throughout the genome may also be important. Untangling the role of these different characteristics in explaining the rate of sex chromosome divergence will require very large-scale comparative datasets and phylogenetic methods. Work in this direction has started[104], but much more work is needed.

References

1. Bachtrog, D. *et al.* Are all sex chromosomes created equal? *Trends Genet.* **27,** 350–357 (2011).
2. Bachtrog, D. *et al.* Sex determination: why so many ways of doing it? *PLoS Biol.* **12,** e1001899 (2014).
3. Beukeboon, L. W. & Perrin, N. *The Evolution of Sex Determination* (Oxford University Press, 2014).
4. Bergero, R. & Charlesworth, D. The evolution of restricted recombination in sex chromosomes. *Trends Ecol. Evol.* **24,** 94–102 (2009).
5. Muller, H. Genetic variability, twin hybrids and constant hybrids, in a case of balanced lethal factors. *Genetics* **3,** 422–499 (1914).
6. Bachtrog, D. Adaptation shapes patterns of genome evolution on sexual and asexual chromosomes in *Drosophila*. *Nat. Genet.* **34,** 215–219 (2003).
7. Bachtrog, D. Y chromosome evolution: emerging insights into processes of Y chromosome degeneration. *Nat. Rev. Genet.* **14,** 113–124 (2013).
8. Vicoso, B. & Charlesworth, B. Evolution on the X chromosome: unusual patterns and processes. *Nat. Rev. Genet.* **7,** 645–653 (2006).
9. Lahn, B. T. & Page, D. C. Four evolutionary strata on the human X chromosome. *Science* **286,** 964–967 (1999).
   Evidence of strata on the human sex chromosomes may indicate a role for inversions in sex chromosome divergence.
10. Skaletsky, H. *et al.* The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423,** 825–837 (2003).
11. Soh, Y. Q. S. *et al.* Sequencing the mouse Y chromosome reveals convergent gene acquisition and amplification on both sex chromosomes. *Cell* **159,** 800–813 (2014).
12. Charlesworth, B. Model for the evolution of Y chromosomes and dosage compensation. *Proc. Natl Acad. Sci. USA* **75,** 5618–5622 (1978).
13. Rice, W. R. The accumulation of sexually antagonistic genes as a selective agent promoting the evolution of reduced recombination between primitive sex chromosomes. *Evolution* **41,** 911–914 (1987).
14. Rice, W. R. Evolution of the sex chromosome in animals. *Bioscience* **46,** 331–343 (1996).
15. Charlesworth, B. The evolution of sex chromosomes. *Science* **251,** 1130–1133 (1990).
16. Bachtrog, D. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. *Nat. Rev. Genet.* **14,** 113–124 (2013).
17. Kamiya, T. *et al.* A trans-species missense SNP in Amhr2 is associated with sex determination in the tiger pufferfish, *Takifugu rubripes* (*Fugu*). *PLoS Genet.* **8,** e1002798 (2012).
    Genetic differences between sex chromosomes can be as simple as a single SNP.
18. Mank, J. E. & Ellegren, H. Parallel divergence and degradation of the avian W sex chromosome. *Trends Ecol.* **22,** 389–391 (2007).
19. Vicoso, B., Kaiser, V. B. & Bachtrog, D. Sex-biased gene expression at homomorphic sex chromosomes in emus and its implications for sex chromosome evolution. *Proc. Natl Acad. Sci. USA* **110,** 6453–6458 (2013).
20. Vicoso, B., Emerson, J. J., Zektser, Y., Manajan, S. & Bachtrog, D. Comparative sex chromosome divergence in snakes: differentiation and lack of global dosage compensation. *PLoS Biol.* **11,** e1001643 (2013).
21. Stock, M. *et al.* Ever-young sex chromosomes in European tree frogs. *PLoS Biol.* **9,** e1001062 (2011).
    Homomorphic sex chromosomes in tree frogs are old, indicating that sex chromosome divergence in not inevitable.

22. Dobzhansky, T. Genetics of natural populations XVIII. Experiments on chromosomes of *Drosophila pseudoobscura* from different geographic regions. *Genetics* **33,** 588–602 (1948).
23. Dobzhansky, T. *Genetics of the Evolutionary Process* (Columbia University Press, 1970).
24. Dobzhansky, T. & Pavlovsky, O. Indeterminate outcome of certain experiments on Drosophila populations. *Evolution* **7,** 198–210 (1953).
25. Dobzhansky, T. & Pavlovsky, O. Interracial hybridization and breakdown of coadapted gene complexes in *Drosophila paulistorum* and *Drosophila willistoni*. *Proc. Natl Acad. Sci. USA* **44,** 622–629 (1958).
26. Joron, M. *et al.* Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature* **477,** 203–205 (2011).
27. Kupper, C. *et al.* A supergene determines highly divergent male reproductive morphs in the ruff. *Nat. Genet.* **48,** 79–83 (2016).
28. Lamichhaney, S. *et al.* Structural genomic changes underlie alternative reproductive strategies in the ruff (*Philomachus pugnax*). *Nat. Genet.* **48,** 84–88 (2016).
29. Wang, J. *et al.* A Y-like social chromosome causes alternative colony organization in fire ants. *Nature* **493,** 664–668 (2013).
Supergenes underlying complex phenotypic variation can mimic sex chromosomes.
30. Feder, J. L. & Nosil, P. The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. *Evolution* **64,** 1729–1747 (2010).
31. Bull, J. J. *Evolution of Sex Determining Mechanisms* (Benjamin Cummings, 1983).
32. Fisher, R. A. The evolution of dominance. *Biol. Rev.* **6,** 345–368 (1931).
33. Winge, Ö The location of eighteen genes in *Lebistes reticulata*. *J. Genet.* **18,** 1–43 (1927).
Early work indicating that many colour genes are Y linked in the guppy set the stage for current theories about the role of sexual conflict in sex chromosome evolution.
34. Gorelick, R. Evolution of dioecy and sex chromosomes via methylation driving Muller's ratchet. *Biol. J. Linn. Soc.* **80,** 353–368 (2003).
35. Ironside, J. No amicable divorce? Challenging the notion that sexual antagonism drives sex chromosome evolution. *Bioessays* **32,** 718–726 (2010).
36. Lange, J. *et al.* Isodicentric Y chromosomes and sex disorders as byproducts of homologous recombination that maintains palindromes. *Cell* **138,** 855–869 (2009).
37. Chippindale, A. K. & Rice, W. R. Y chromosome polymorphism is a strong determinant of male fitness in *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA* **98,** 5677–5682 (2001).
38. Lemos, B., Araripe, L. O. & Hartl, D. L. Polymorphic Y chromosomes harbor cryptic variation with manifold functional consequences. *Science* **319,** 91–93 (2008).
39. Moghadam, H. K., Pointer, M. A., Wright, A. E., Berlin, S. & Mank, J. E. W chromosome expression responds to female-specific selection. *Proc. Natl Acad. Sci. USA* **109,** 8207–8211 (2012).
40. Koerich, L. B., Wang, X. Y., Clark, A. G. & Carvalho, A. B. Low conservation of gene content in the *Drosophila* Y chromosome. *Nature* **456,** 949–951 (2008).
41. Zhou, Q. & Bachtrog, D. Sex-specific adaptation drives early sex chromosome evolution in *Drosophila*. *Science* **337,** 341–345 (2012).
42. Gamble, T., Geneva, A. J., Glor, R. E. & Zarkower, D. *Anolis* sex chromosomes are derived from a single ancestral pair. *Evolution* **68,** 1027–1041 (2014).
43. Rovatsos, M., Altmanova, M., Pokorna, M. & Kratochvil, L. Conserved sex chromosomes across adaptively radiated *Anolis* lizards. *Evolution* **68,** 2079–2085 (2014).
44. Miura, I., Ohtani, H., Nakamura, M., Ichikawa, Y. & Saitoh, K. The origin and differentiation of the heteromorphic sex chromosomes Z, W, X, and Y in the frog *Rana rugosa*, inferred from the sequences of a sex-linked gene, ADP/ATP translocase. *Mol. Biol. Evol.* **15,** 1612–1619 (1998).
45. Yoshimoto, S. *et al.* A W-linked DM-domain gene, DM-W, participates in primary ovary development in *Xenopus laevis*. *Proc. Natl Acad. Sci. USA* **105,** 2469–2474 (2008).
46. Matsubara, K. *et al.* Evidence for different origin of sex chromosomes in snakes, birds, and mammals and step-wise differentiation of snake sex chromosomes. *Proc. Natl Acad. Sci. USA* **103,** 18190–18195 (2006).
47. Mank, J. E., Promislow, D. E. L. & Avise, J. C. Evolution of alternative sex-determining mechanisms in teleost fishes. *Biol. J. Linn. Soc.* **87,** 83–93 (2006).
48. Hough, J., Hollister, J. D., Wang, W., Barrett, S. C. H. & Wright, S. I. Genetic degeneration of old and young Y chromosomes in the flowering plant *Rumex hastatulus*. *Proc. Natl Acad. Sci. USA* **111,** 7713–7718 (2014).
49. Liu, Z. Y. *et al.* A primitive Y chromosome in papaya marks incipient sex chromosome evolution. *Nature* **427,** 348–352 (2004).
50. Papadopulos, A. S. T., Chester, M., Ridout, K. & Filatov, D. A. Rapid Y degeneration and dosage compensation in plant sex chromosomes. *Proc. Natl Acad. Sci. USA* **112,** 13021–13026 (2015).
51. Spigler, R. B., Lewers, K. S., Main, D. S. & Ashman, T. L. Genetic mapping of sex determination in a wild strawberry, *Fragaria virginiana*, reveals earliest form of sex chromosome. *Heredity* **101,** 507–517 (2008).
52. Kitano, J. *et al.* A role for a neo-sex chromosome in stickleback speciation. *Nature* **461,** 1079–1083 (2009).
53. Natri, H. M., Shikano, T. & Merila, J. Progressive recombination suppression and differentiation in recently evolved neo-sex chromosomes. *Mol. Biol. Evol.* **30,** 1131–1144 (2013).
54. Roberts, R. B., Ser, J. R. & Kocher, T. D. Sexual conflict resolved by invasion of a novel sex determiner in Lake Malawi cichlid fishes. *Science* **326,** 998–1001 (2009).
55. Ser, J. R., Roberts, R. B. & Kocher, T. D. Multiple interacting loci control sex determination in Lake Malawi cichlid fish. *Evolution* **64,** 486–501 (2010).
56. Renner, S. S. The relative and absolute frequencies of angiosperm sexual systems: Dioecy, monoecy, gynodioecy and an updated online database. *Am. J. Bot.* **101,** 1588–1596 (2014).
57. Ming, R., Bendahmane, A. & Renner, S. S. Sex chromosomes in land plants. *Annu. Rev. Plant. Biol.* **62,** 485–514 (2011).
58. Charlesworth, B. & Charlesworth, D. Model for evolution of dioecy and gynodioecy. *Am. Nat.* **112,** 975–997 (1978).
59. Tennessen, J. A., Govindarajulu, R., Liston, A. & Ashman, T. L. Targeted sequence capture provides insight into genome structure and genetics of male sterility in a gynodioecious diploid strawberry *Fragaria vesca* ssp. *bracteata* (Rosaceae). *G3 (Bethesda)* **3,** 1341–1351 (2013).
60. Wang, J. P. *et al.* Sequencing papaya X and Y-h chromosomes reveals molecular basis of incipient sex chromosome evolution. *Proc. Natl Acad. Sci. USA* **109,** 13710–13715 (2012).
61. Bell, G. *The Masterpiece of Nature: The Evolution and Genetics of Sexuality* (University of California Press, 1982).
62. Lenormand, T. The evolution of sex dimorphism in recombination. *Genetics* **163,** 811–822 (2003).
63. Lenormand, T. & Dutheil, J. Recombination difference between sexes: A role for haploid selection. *PLoS Biol.* **3,** 396–403 (2005).
The theoretical predictions about sex differences in recombination may be important in understanding early stages of sex chromosome evolution.
64. Vicoso, B. & Bachtrog, D. Numerous transitions of sex chromosomes in Diptera. *PLoS Biol.* **13 (2015).**
65. Charlesworth, D., Charlesworth, B. & Marais, G. Steps in the evolution of heteromorphic sex chromosomes. *Heredity* **95,** 118–128 (2005).
66. Bergero, R., Forrest, A., Kamau, E. & Charlesworth, D. Evolutionary strata on the X chromosomes of the dioecious plant *Silene latifolia*: Evidence from new sex-linked genes. *Genetics* **175,** 1945–1954 (2007).
67. Roesti, M., Moser, D. & Berner, D. Recombiantion in the threespine stickleback genome—patterns and consequences. *Mol. Ecol.* **22,** 3014–3027 (2013).
68. Wright, A. E., Moghadam, H. K. & Mank, J. E. Trade-off between selection for dosage compensation and masculinization on the avian Z chromosome. *Genetics* **192,** 1433–1445 (2012).
69. Bergero, R., Qiu, S., Forrest, A., Borthwick, H. & Charlesworth, D. Expansion of the pseudo-autosomal region and ongoing recombination suppression in the *Silene latifolia* sex chromosomes. *Genetics* **194,** 673–686 (2013).
70. Chibalina, M. V. & Filatov, D. A. Plant Y chromosome degeneration is retarded by haploid purifying selection. *Curr. Biol.* **21,** 1475–1479 (2011).
Systems with strong haploid selection may exhibit slow rates of sex chromosome divergence.
71. Kong, A. *et al.* Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467,** 1099–1103 (2010).
Recombination hotspots can vary substnatially between the sexes, and this many be important in sex chromosome formation.
72. Flot, J. F. *et al.* Genomic evidence for ameiotic evolution in the bdelloid rotifer *Adineta vaga*. *Nature* **500,** 453–457 (2013).
73. Baudat, F. *et al.* PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* **327,** 836–840 (2010).
74. Berg, I. L. *et al.* Variants of the protein PRDM9 differentially regulate a set of human meiotic recombination hotspots highly active in African populations. *Proc. Natl Acad. Sci. USA* **108,** 12378–12383 (2011).
75. Parvanov, E. D., Petkov, P. M. & Paigen, K. *Prdm9* controls activation of mammalian recombination hotspots. *Science* **327,** 835–835 (2010).
76. Auton, A. *et al.* Genetic recombination is targeted towards gene promoter regions in dogs. *PLoS Genet.* **9,** e1003984 (2013).
77. Singhal, S. *et al.* Stable recombination hotspots in birds. *Science* **350,** 928–932 (2015).
78. Auton, A. *et al.* A fine-scale chimpanzee genetic map from population sequencing. *Science* **336,** 193–198 (2012).
79. Myers, S. *et al.* Drive against hotspot motifs in primates implicates the *PRDM9* gene in meiotic recombination. *Science* **327,** 876–879 (2010).

80. Hinch, A. G. *et al.* The landscape of recombination in African Americans. *Nature* **476,** 170–175 (2011).
81. Choi, K. & Henderson, I. R. Meiotic recombination hotspots - a comparative view. *Plant J.* **83,** 52–61 (2015).
82. Charlesworth, D. Plant sex chromosome evolution. *J. Exp. Bot.* **64,** 405–420 (2013).
83. Ahmed, S. *et al.* A haploid system of sex determination in the brown alga *Ectocarpus sp. Curr. Biol.* **24,** 1945–1957 (2014).
84. Bergero, R., Qiu, S. & Charlesworth, D. Gene loss from a plant sex chromosome system. *Curr. Biol.* **25,** 1234–1240 (2015).
85. Charlesworth, B. Model for the evolution of Y chromosomes and dosage compensation. *Proc. Natl Acad. Sci. USA* **75,** 5618–5622 (1978).
86. Ohno, S. *Sex Chromosomes and Sex Linked Genes* (Springer-Verlag, 1967).
87. Mank, J. E. Sex chromosome dosage compensation: definitely not for everyone. *Trends Genet.* **29,** 677–683 (2013). ***Dosage compensation may be important in the rate of sex chromosome divergence*.**
88. Mank, J. E. The W, X Y and Z of sex-chromosome dosage compensation. *Trends Genet.* **25,** 226–233 (2009).
89. Mullon, C., Wright, A. E., Reuter, M., Pomiakowski, A. & Mank, J. E. Evolution of dosage compensation under sexual selection diffes between X and Z chromosomes. *Nat. Commun.* **6,** 7720 (2015).
90. Payer, B. & Lee, J. T. X chromosome dosage compensation: How mammals keep the balance. *Annu. Rev. Genet.* **42,** 733–772 (2008).
91. Zhou, Q. *et al.* The epigenome of evolving *Drosophila* neo-sex chromosomes: Dosage compensation and heterochromatin formation. *PLoS Biol.* **11,** e1001711 (2013).
92. Zhou, Q. *et al.* Complex evolutionary trajectories of sex chromosomes across bird taxa. *Science* **346,** 1246338 (2014).
93. Naurin, S., Hansson, B., Bensch, S. & Hassequist, D. Why does dosage compensation differ between XY and ZW taxa? *Trends Genet.* **26,** 15–20 (2010).
94. White, M., Kitano, J. & Peichel, C. L. Purifying selection maintains dosage sensitive genes during degeneration of the threespine stickleback Y chromosome. *Mol. Biol. Evol.* **32,** 1981–1995 (2015).
95. Nakamura, M. Sex determination in amphibians. *Semin. Cell Dev. Biol.* **20,** 271–282 (2009).
96. Wallace, H., Badawy, G. M. I. & Wallace, B. M. N. Amphibian sex determination and sex reversal. *Cell. Mol. Life Sci.* **55,** 901–909 (1999).
97. McNair, A., Lokman, P. M., Closs, G. P. & Nakagawa, S. Ecological and evolutionary applications for environmetnal sex reversal of fish. *Q. Rev. Biol.* **90,** 23–44 (2015).
98. Perrin, N. Sex reversal: a fountain of youth for sex chromosomes? *Evolution* **63,** 3043–3049 (2009).
99. Dufresnes, C. *et al.* Sex-chromosome homomorphy in palearctic tree frogs results from both turnovers and X-Y recombination. *Mol. Biol. Evol.* **32,** 2328–2337 (2015).
100. Stock, M. *et al.* Low rates of X-Y recombination, not turnovers, account for homomorphic sex chromosomes in several diploid species of Palearctic green toads (*Bufo viridis* subgroup). *J. Evol. Biol.* **26,** 674–682 (2013).
101. Charlesworth, D. & Mank, J. E. The birds and the bees and the flowers and the trees: lessons from genetic mapping of sex determination in plants and animals. *Genetics* **186,** 9–31 (2010).
102. van Doorn, G. S. & Kirkpatrick, M. Turnover of sex chromosomes induced by sexual conflict. *Nature* **449,** 909–912 (2007).
103. van Doorn, G. S. & Kirkpatrick, M. Transitions between male and female heterogamety caused by sex-antagonistic selection. *Genetics* **186,** 629–645 (2010).
104. Tree of Sex Consortium. Tree of sex consortium: a database of sexual systems. *Sci. Data* **1,** 140015 (2014).
105. Ross, M. T. *et al.* The DNA sequence of the human X chromosome. *Nature* **434,** 325–337 (2005).
106. Lemaitre, C. *et al.* Footprints of inversions at present and past pseudoautosomal boundaries in human sex chromosomes. *Genome Biol. Evol.* **1,** 56–66 (2009).
107. Wilson, M. A. & Makova, K. D. Evolution and survival on eutherian sex chromosomes. *PLoS Genet.* **5,** e1000568 (2009).
108. Pandey, R. S., Sayres, M. A. W. & Azad, R. K. Detecting evolutionary strata on the human X chromosome in the absence of gametologous Y-linked sequences. *Genome Biol. Evol.* **5,** 1863–1871 (2013).
109. Wright, A. E., Harrison, P. W., Montgomery, S. H., Pointer, M. A. & Mank, J. E. Independent stratum formation on the avian sex chromosomes reveals inter-chromosomal gene conversion and predominance of purifying selection on the W chromosome. *Evolution* **68,** 3281–3295 (2014).
110. Morgan, T. Complete linkage in the second chromosome of male *Drosophila*. *Science* **36,** 719–720 (1912).
111. Haldane, J. Sex ratio and unisexual sterility in hybrid animals. *J. Genet.* **12,** 101–109 (1922).
112. Suomalai, E., Cook, L. M. & Turner, J. R. G. Achiasmatic oogenesis in Heliconiine butterflies. *Hereditas* **74,** 302–304 (1973).
113. Nokkala, S. & Nokkala, C. Achiasmatic male meiosis in two species of *Saldula* (Salidae, Hemiptera). *Hereditas* **99,** 131–134 (1983).
114. Bardella, V. B., Gil-Santana, H. R., Panzera, F. & Vanzela, A. L. L. Karyotype diversity among predatory Reduviidae (Heteroptera). *Comp. Cytogenet.* **8,** 351–367 (2014).
115. Grozeva, S. & Nokkala, S. Chromosomes and their meiotic behavior in two families of the primitive infraorder Dipsocoromorpha (Heteroptera). *Hereditas* **125,** 31–36 (1996).
116. Poggio, M. G., Di Iorio, O., Turienzo, P., Papeschi, A. G. & Bressa, M. J. Heterochromatin characterization and ribosomal gene location in two monotypic genera of bloodsucker bugs (Cimicidae, Heteroptera) with holokinetic chromosomes and achiasmatic male meiosis. *Bull. Entomol. Res.* **104,** 788–793 (2014).
117. White, M. J. D. Chiasmatic and achiasmatic meiosis in African eumastacid grasshoppers. *Chromosoma* **16,** 271–307 (1965).
118. Brooks, L. D. & Marks, R. W. The organization of genetic varaition for recombination in *Drosophila melanogaster*. *Genetics* **114,** 525–547 (1986).
119. Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310,** 321–324 (2005).
120. Burt, A., Bell, G. & Harvey, P. H. Sex-differences in recombination. *J. Evol. Biol.* **4,** 259–277 (1991).
121. Wyman, M. J. & Wyman, M. C. specific recombination rates and allele frequencies affect the invasion of sexually antagonistic variation on autosomes. *J. Evol. Biol.* **26,** 2428–2437 (2013).

Acknowledgements

This review was originally inspired during JEM's fellowship at the Wissenschaftskolleg zu Berlin. We gratefully acknowledge support from the European Research Council (Grant Agreements 260233 and 680951 to J.E.M.), the Australian Research Council (DE150101853 to R.D.) and Marie Curie Actions (Grant Agreement 655392 to R.D.). We thank Sofia Berlin, Lynda Delph and John Pannell for helpful discussions.

Author contributions

All authors wrote the manuscript.

Additional information

Competing financial interest: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Wright, A.E. *et al.* How to make a sex chromosome. *Nat. Commun.* 7:12087 doi: 10.1038/ncomms12087 (2016).

The following article was first published in

Nucleic Acids Research

Schmitz JS, **Zimmer F**, Bornberg-Bauer E (2016) Mechanisms of transcription factor evolution in Metazoa. *Nucleic Acids Research* 13: 6287-6297
doi: 10.1093/nar/gkw492

Published online 10 June 2016 *Nucleic Acids Research, 2016, Vol. 44, No. 13* ***6287–6297***
doi: 10.1093/nar/gkw492

Mechanisms of transcription factor evolution in Metazoa

Jonathan F. Schmitz[1], **Fabian Zimmer**[1,2] **and Erich Bornberg-Bauer**[1,*]

[1]Evolutionary Bioinformatics Group, Institute for Evolution and Biodiversity, Hüfferstrasse 1, D-48149 Münster, Germany and [2]Department of Genetics, Evolution and Environment, University College London, London WC1E 6BT, UK

Received January 28, 2016; Revised May 18, 2016; Accepted May 22, 2016

ABSTRACT

Transcriptions factors (TFs) are pivotal for the regulation of virtually all cellular processes, including growth and development. Expansions of TF families are causally linked to increases in organismal complexity. Here we study the evolutionary dynamics, genetic causes and functional implications of the five largest metazoan TF families. We find that family expansions dominate across the whole metazoan tree; however, some branches experience exceptional family-specific accelerated expansions. Additionally, we find that such expansions are often predated by modular domain rearrangements, which spur the expansion of a new sub-family by separating it from the rest of the TF family in terms of protein–protein interactions. This separation allows for radical shifts in the functional spectrum of a duplicated TF. We also find functional differentiation inside TF sub-families as changes in expression specificity. Furthermore, accelerated family expansions are facilitated by repeats of sequence motifs such as C2H2 zinc fingers. We quantify whole genome duplications and single gene duplications as sources of TF family expansions, implying that some, but not all, TF duplicates are preferentially retained. We conclude that trans-regulatory changes (domain rearrangements) are instrumental for fundamental functional innovations, that cis-regulatory changes (affecting expression) accomplish wide-spread fine tuning and both jointly contribute to the functional diversification of TFs.

INTRODUCTION

Transcriptional regulation is crucial for all known processes in life, in particular for growth and development. Consequently, the evolution of gene expression regulation is tightly linked to the apparent evolution of biological complexity, for example as measured in the number of cell types (1–7). The underlying genomic changes are, as yet, only poorly understood but, among others, changes in cis-regulatory elements (8,9), transcription associated proteins (10) and small regulatory RNAs (11,12) have been identified as major contributors to genomic adaptation. Transcription factors (TFs) are proteins which regulate the transcription of DNA to mRNA in all known organisms by binding to specific DNA target sequences. In eukaryotes, TFs play an important role in development, cellular organization and signal response (13) and dis- or non-functional TF genes have been linked to a number of diseases such as cancer (14,15). TFs have also been implicated in evolutionary innovation of novel phenotypes and developmental frameworks (16).

Generally, expansions of gene families involved in signaling and regulation can be observed at a much higher frequency than the expansions of, e.g. metabolic pathways (17). Also, the number of TFs per genome was found to correlate over-proportionally with the number of genes in genomes, resulting in a higher proportion of TF genes in larger genomes (18). This high proportion of TFs in large genomes suggests that higher complexity requires an over-proportional increase in regulatory elements.

The increases in the number of regulatory proteins in general (4,19) and of TFs in particular (5) have repeatedly been connected to phenotypic innovations and the evolution of more complex organisms. For example, TF family expansions (and size reductions) have been implicated in emergence (and loss) of complex features in Stramenopiles (20) and Viridiplantae (21). Another recent example is the expansion of the C2H2 zinc finger (ZF) and the protocadherin families, which has been linked to increased morphological and developmental complexity of the octopus (22). An expansion of the C2H2 ZF TF family has also been linked to the the secondarily evolved multicellularity in the red algae Chondrus (21,23).

Furthermore, the emergence of new TFs has also been shown to play a role in phenotypic changes, especially in animals (24). Taken together, these and other findings suggest that emergence of TFs and growth of TF families are both

[*]To whom correspondence should be addressed. Tel: +49 251 83 21630; Fax: +49 251 83 24668; Email: ebb@uni-muenster.de

related to increases in morphological complexity and the number of cell types (2,3,20,21,25,26). Therefore, a detailed cross-species comparison of TF repertoires is important to delineate which genetic events underlie the expansion of TF families and which ones were instrumental in creating fundamentally new phenotypes which have led to new and possibly more complex body plans. Indeed, with the availability of many genomes such large scale comparisons allow a detailed analysis of origin and nature of important molecular changes in TFs.

On larger evolutionary time scales, the emergence of new TFs or TF sub-families has been linked to many major transitions in morphology and development, e.g. to the emergence of multicellularity (2,25) or the emergence of flowering plants (27). Indeed, most of the largest metazoan TF families originated already before the emergence of Metazoa and thus multicellularity (25). The further expansion of these families then allowed for the evolution of increasingly complex organisms in Metazoa (2).

In some TF families the expansion results clearly from a number of single gene duplications (SGDs) (28). On the other hand, some expansions were suspected to have been triggered mainly by whole genome-duplications (WGDs), coupled with a high retention rate of TFs (3,5,25,29,30). However, it is still unclear if and how WGDs are instrumental in supporting higher regulatory and organismal complexity. First, no consensus has been reached regarding the causes of the high retention rate of TF genes after SGD as well as WGD events (31). Second, WGDs could not be linked to increased complexity as it is documented in the metazoan fossil record (32). Nevertheless, both processes (SGD and WGD) seem to play a role in the expansion of gene families, specifically TF families (31).

It has been proposed that the number of TF family members is limited by the number of possible target sequences (33). These findings imply that dimerization of TFs would allow for TF family expansion by doubling the DNA target sequence length (one target sequence for both proteins in the dimer). Indeed, many TF families form protein dimers or larger protein complexes in order to bind DNA. Within these TF families, some members are only able to form complexes with themselves (homodimerize), while others can dimerize with other members of the family (heterodimerize) (28,34,35). The interactions between TF members form large interaction networks and the structure of these networks depends on the TF family (35). However, most of the interactions in complex formation are context dependent, i.e. preference may change depending on e.g. pH, localization, concentration or salt strength (36,37). This volatility induces a highly entangled combinatorial interaction pattern which helps to increase the capacity for regulatory fine-tuning, way beyond the associated increase in the number of TFs.

Because several hundred millions of years have elapsed since the emergence of most TF families, it is only rarely possible (see e.g. (38)) to track down the precise molecular and genetic origin of new TF families. Nonetheless, in many cases comparative genomics can reveal major rearrangements which shifted functions of TFs and triggered the emergence of new sub-families. For example, the loss and gain of additional domains has been reported in several families of TFs (35,39). Such changes often entail a strongly altered functional spectrum by changing binding specificites to DNA and upstream regulatory proteins, e.g. signaling proteins or other transcriptional regulators (21,28,40). Domain rearrangements (DRs) may thus explain 'functional shifts', i.e. sudden, radical changes in the regulatory potential of TFs.

In this study we ask how strong the effects of WGDs, SGDs and DRs are on the growth of TF families. Additionally, we analyze if any of these genomic events, or a combination of them, have led to functional shifts which my have spurred fundamental developmental innovations. Accordingly, we study the evolution of the five largest TF families (26) and the p53 family in 36 metazoan species to elucidate the evolutionary history of these families during the evolution of more complex, multicellular organisms. The selected genomes and the size of the chosen families provide a relatively dense and even distribution across the metazoan tree along which many complex phenotypes evolved. We determine extant and ancestral TF family sizes to identify branches with accelerated expansions and relate expansions to underlying molecular changes and genomic rearrangements. Finally, we relate these changes to functional properties which can be inferred from annotations and expression profiles of TFs.

MATERIALS AND METHODS

Taxon sampling and sequence data

The 36 species analyzed here were selected to represent a large sample of sequenced Metazoa with a high quality genome available. *Saccharomyces cerevisiae* was chosen as a non-metazoan outgroup with a high genome quality. To enable phylogenetic analyses, a dated tree was reconstructed based on the study by Erwin *et al.* (41). Dating for species not included in the Erwin *et al.* study were added manually according to various sources (see Supplementary Data). The sequence data for most species were obtained from Ensembl release 74 (42) or from Ensembl Genomes release 21 (43). Species not available on Ensembl were downloaded from various sources, see Supplementary Table S3. Only the longest splicing variant of each gene was considered in our analyses.

Domain annotation

Domains were annotated using the hidden Markov models (HMMs) of Pfam-A version 27.0 (44). The PfamScan script provided by Pfam was used to perform the annotation. A list of HMMs representing the TF families' DNA-binding domains (DBDs) was used to identify TF proteins. For the list defining the relationship DBD–HMMs see Supplementary Table S2. A protein's domain arrangement was defined as the sequence of domains, domain repetitions were not collapsed. All proteins sharing a domain arrangement were grouped into a domain arrangement cluster (DAC).

Ancestral family size reconstruction

Ancestral TF family sizes for all nodes in the species tree were reconstructed using Count (45) in symmetrical Wagner parsimony mode by setting the ratio of gain- to loss-penalties to 1. In Count, the DACs were used as subfamilies. The number of annotated proteins per DAC was used as input for Count.

Comparison of gene/DAC gain/loss rates. For each of the branches of the species tree, the gene/DAC gain and loss rates were calculated by dividing the number of events per category by the branch length in million years. This analysis was performed using a custom R script (46). Figures comparing the distribution of rates were produced using the ggplot2 R library (47). To test the rate distribution of the four categories for differences, the Wilcoxon signed rank test was used. The wilcox.test function of the R base package was used for this purpose (46).

Plotting of TF family evolution per lineage. The TF family evolution of a lineage was represented by plotting the TF family size and DAC composition for each ancestral node. The plotting was performed using a custom Python script utilizing the matplotlib plotting library (48).

Gene Ontology enrichment testing

Gene Ontology (GO) annotation data were downloaded from Ensembl for the model organisms *Homo sapiens*, *Danio rerio*, *Drosophila melanogaster* and *Caenorhabditis elegans* (42). Using the topGO R library (49), the proteins of each DAC were tested for GO enrichment using all proteins of the respective TF family as background. topGO's weighted Fisher test method was used. The minimum number of annotations per GO term was set to 3 (Node Size = 3) to ensure a certain stability of the GO annotations. Consequently, only DACs with at least three protein members were taken into account for this analysis. A *P*-value cutoff of 0.05 was chosen to select only significant hits. A multiple testing correction was performed by multiplying *P*-values with the number of DACs for which GO enrichment tests were executed. In this analysis, only the biological process class of GO was considered.

Gene expression pattern comparison

Expression data for eight human organs (50) were used to compare the expression of the TFs. Pre-computed FPKM values for this experiment were obtained from the Expression Atlas Website (51). To compare the expression patterns among the TF genes, the genes were clustered according to their expression profile similarity using the cosine function as a similarity measurement. Clusters of genes with similar expression profiles were then manually inspected for the proteins' domain arrangements. The vector of expression strengths per organ, given as the FPKM value, was used as expression profile for each gene. This approach was chosen since FPKM values can not be used to reliably compare expression strength across experiments (52). The analysis was conducted in R (46) utilizing the lsa packages' cosine function (53) and hclust in complete mode from the R base package. Custom python scripts were used to analyze expression breadth using a cutoff of 1 FPKM for presence of expression. The first node with DAC presence generated by Count (see above) was used to determine domain arrangement age. GOATOOLS (https://github.com/tanghaibao/Goatools) in Fisher's exact test mode was used to determine GO enrichment in clusters of genes with similar expression patterns. Clusters of genes with similar expression were extracted using the hierarchical clustering function of SciPy (54).

RESULTS AND DISCUSSION

We annotated TFs in 36 metazoan species using HMMs to find the TF family-specific DBDs. We leave aside other transcriptional regulators (see also (10)), because most of these have multiple, more general, roles such that their evolutionary functional impacts are even more difficult to characterize than those of TFs. We do, however, include family members of TFs that have lost their DNA binding abilities in a secondary event. Such a loss of DNA binding affinity can provide valuable information on the molecular triggers of functional shifts and family expansions and can be clearly delineated by comparative genomics.

To analyze TF family sizes, we first determined the TF families in our set of 36 metazoan species and baker's yeast (see Figure 1). The first family we annotate is the bHLH TF family, which is characterized by the basic helix-loop-helix domain, in which the basic region binds DNA and the helix-loop-helix motif facilitates dimerization and DNA binding. Next to the bHLH domain, other protein domains can be found in bHLH proteins (55), such as the Orange, PAS or Leucine zipper (LZ) domains (28). These domains can have various functions, such as environmental sensing, signal transduction and dimerization facilitation (56,57).

The second family is the bZIP TF family, whose proteins contain a basic region that binds DNA, just as bHLH proteins do. However, the bZIP basic region does not show any detectable homology to the bHLH basic region and is likely an example of convergent evolution. In bZIP proteins the basic region directly extends into an LZ which, convergently to bHLH proteins again, facilitates dimerization (58). The bZIP family comprises many well known TFs such as JUN and FOS, which are involved in cell proliferation, apoptosis, and survival, as well as cancer development in case of loss-of-function mutations (59,60).

The third family is the Homeobox TF family, which is defined by the Homeobox domain that consists of 60 amino acids forming three α-helices (61). Proteins carrying Homeobox domains can be found in all eukaryotes (25) and play an important role in regulating development, especially in Metazoa (61). The Hox genes are the best-known metazoan homeobox genes and are crucial in Bilateria for determining the body axis during development among other functions (62).

Fourth, the Nuclear Receptor (NR) family was analyzed. NR proteins contain a DBD and a ligand-binding domain (LBD). The LBD binds a number of cofactors such as steroid hormones or lipids (63,64) and can also facilitate dimerization (65). The NR family is Metazoa-specific (25) and important for the regulating of development, metabolism and reproduction.

Next, the C2H2 ZF family is defined by a sequence motif in which two Cystein (C) and two Histidine (H) amino acid residues coordinate a zinc ion. The C2H2 ZF domain fa-

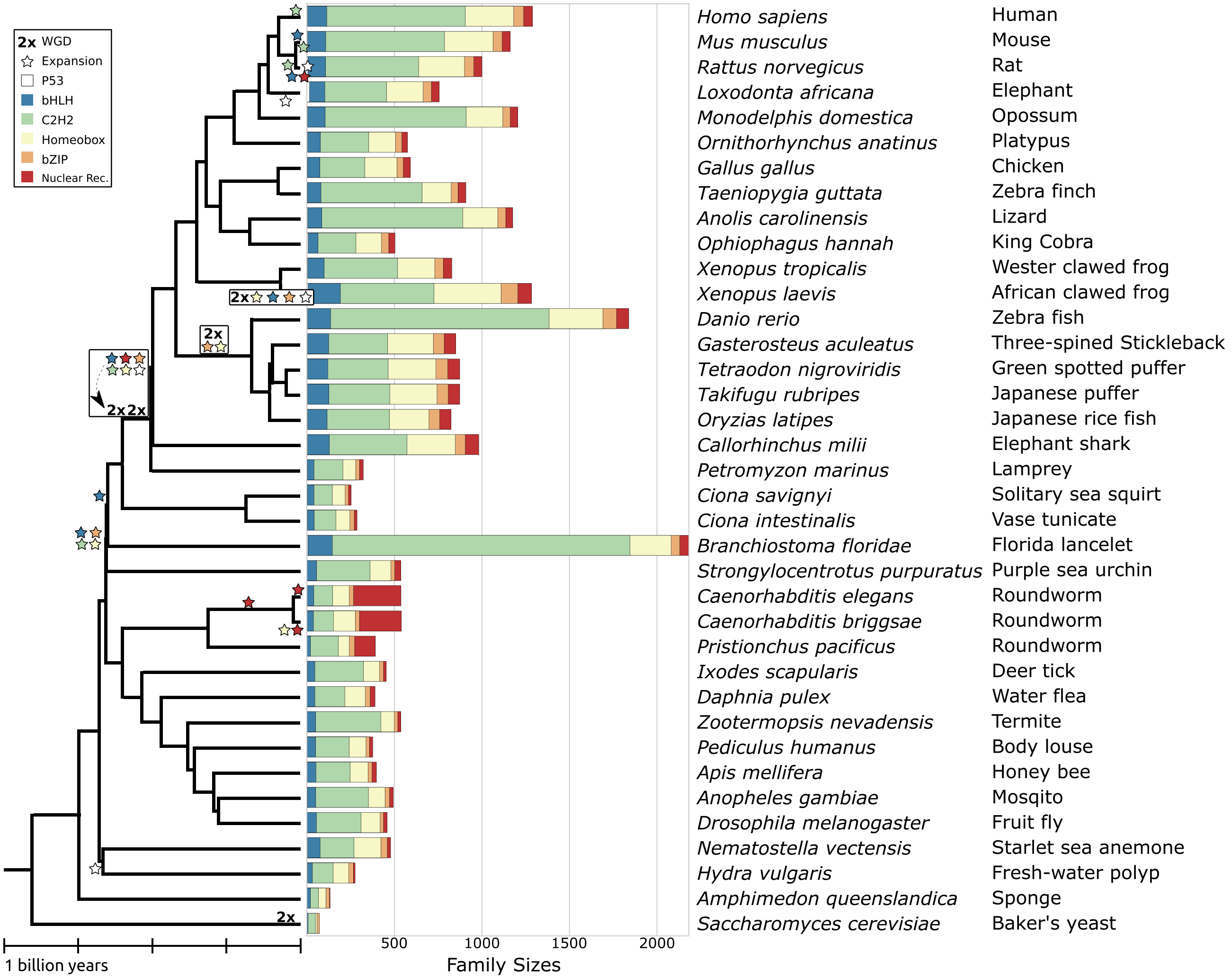


Figure 1. Stacked bar plots depicting TF family sizes in analyzed species. The left-hand side of the graph shows a phylogenetic tree of the analyzed species. WGD events are denoted with a '2x'-symbol. Branches with accelerated gene gain rates are highlighted with stars. The stars are colored according to the TF family with accelerated gain rate. Time scale is approximate and largely based on (41). See 'Materials and Methods' section for more details.

cilitates DNA binding as well as dimerization and is made up of two β-sheets and one α-helix. C2H2 ZF genes can be found in all eukaryotes (25) and have various functions such as regulation of stress response (66). C2H2 ZFs have been proposed to play a role in a number of important evolutionary processes such as speciation in the primate lineage (67,68).

Finally, p53 proteins consist of the p53 DBD, a 200 amino acids long domain consisting mainly of β-sheets, the p53 tetramerization domain that facilitates oligomerization of p53 proteins and in some cases additional domains (69). p53 genes can be found in Holozoa and are not Metazoa-specific (25). In Metazoa, p53 proteins are important, mainly in controlling the cell cycle. Loss of function of a p53 gene can entail a cancer risk (70). The p53 family was included because of this high relevance for medical issues. The other families were analyzed because they are the largest TF families in human and as such represent the bulk of the TFs.

TF family sizes in Metazoa show a pattern of repeated expansions

We analyzed the evolution of TF families in Metazoa using the TF family sizes determined in the previous step. TF family sizes vary drastically in Metazoa for different species and TF families. More specifically, some lineages, such as the ray-finned fishes, have experienced expansion of all TF families. Also, all TF families are expanded compared to most non-vertebrate species. Interestingly, the lancelet lineage has much larger TF families than any of the closely related lineages. On the other hand, in some lineages only one specific TF family is expanded, like the nematode clade in which the largest NR families of all Metazoa can be found (already noted in (71)). Many species' genomes contain a markedly larger C2H2 ZF family compared to closely related species. Examples for species with expanded C2H2 ZF family are *Anolis carolinensis*, *D. rerio*, *Anopheles gambiae*, *Zootermopsis nevadensis* and *Ixodes scapularis*. Additionally, most mammals except for the elephant possess larger

C2H2 ZF families than most other vertebrates. Generally, repeating patterns of clade or lineage-specific expansions of one or multiple TF families can be observed.

The differences in TF family sizes between different animals raises the question of which proportion of species' proteomes the TF families take up. A comparison between the TF family sizes and proteome sizes shows that differences in the proportion of TFs between clades and lineages can be observed (Supplementary Figure S2). In many clades with TF family expansions, TF families make up a larger portion of the proteome. One example are vertebrates in which the bHLH and bZIP families make up a larger portion of the proteome than in non-vertebrate species. The C2H2 ZF family forms a different pattern characterized by lineage-specific expansions. Consequently, the C2H2 TF family makes up a noticeably larger portion of some species' proteome compared to closely related species. C2H2 ZFs are noticeably expanded in the proteome of *D. rerio*, *Branchiostoma floridanus* and *A. carolinensis* for example. The C2H2 family is exceptional in showing such taxonomically restricted bursts, resulting in a larger fraction of the species' proteome being made up of the C2H2 family. The p53 family is expanded in the elephant (*Loxodonta africana*) without an expansion of the elephant proteome. This finding confirms previous ones about a p53 family expansion in elephants (72). However, in general, lineage-specific TF expansions should be interpreted cautiously as they can be an artifact of incorrect genome annotations. In general, TF family expansions often lead to a higher proportion of TFs in the proteome. These expansions can be stable in clades, like for the bHLH and bZIP families in vertebrates.

Given the high variability in TF family sizes it can be concluded that TF family expansion/reduction has occurred along many branches of the metazoan tree. However, findings of burst-like TF family expansions the evolution of Metazoa have only been reported for the proto-metazoan stem (2). In cases where a large clade has significantly larger TF families for all TF families (Vertebrata, ray-finned fishes), a clear connection between WGD events on the branches leading to these clades and the TF family expansions can be made. Additionally, the WGD event on the branch leading to *Xenopus laevis* seems to have doubled the size of most TF families except for the C2H2 ZFs. However, in other cases larger TF families can not be linked to WGD events. The lancet *B. floridae*, for example, has a high number of genes for all TF families, but no WGD event has been proposed to have occurred in that lineage. Also, the many cases of significantly larger C2H2 families do not seem to be connected to WGD events, just as the large NR TF family in nematodes. The hypothesis that C2H2 family expansion is more often connected to SGD than to WGD is further supported by smaller median pairwise gene distances in human (Supplementary Figure S4) and the small amount of C2H2 expansion after the *X. laevis* WGD (see Figure 1). To clarify the relationship between TF family expansions and WGD events, we analyzed ancestral TF family sizes in a next step.

Reconstructed ancestral TF family sizes reveal branches with accelerated gene gain

To locate points in the evolution of Metazoa with accelerated TF family expansion, we reconstructed the TF family sizes of the ancestral nodes of our phylogenetic tree. Using the ancestral TF family sizes we compared the gain/loss rates of genes as well as DACs (genes sharing a domain arrangement) along the branches of the phylogenetic tree. The gain or loss of a DAC describes the gain of at least one gene with a certain domain arrangement or respectively the loss of all genes with a certain domain arrangement in a tree node compared to the parental node. Box plots of gain and loss rates for the six TF families (Supplementary Figure S1) show that the analyzed TF families mainly evolve via gene gain. For all families the gene gain rate distribution has a higher median than the other event types (Wilcoxon signed rank test; $P < 0.01$ for all families). The DAC gain rates are also relatively high compared to the loss rates, which complies with DAC gain being linked to gene gain. The loss rates, for DACs as well as genes, are lower than either of the gain rates, showing that gain of genes seems to be the more important process in TF family evolution. This finding indicates a largely constant growth of the TF families. The magnitude of gene gain rates differs between the six TF families. In p53, for example, the maximum observed gene gain rate is below 0.2 genes per million years, while for C2H2 ZF more than 25 gene gains per million years can be observed on the branch leading to *Mus musculus* since the split from the *Rattus norvegicus* branch.

The gene gain rate distributions (Supplementary Figure S1) feature a number of prominent outliers. These outliers indicate branches with strongly accelerated TF family evolution, indicative of events that we call 'bursts'. For outlier branches with such bursts see Table 1 and Figure 1. Many branches show up for more than one TF family burst, for example the branch leading to *X. laevis* or the Gnathostomata branch. In some cases the bursts in gene gain rate can be linked to WGD events. For the branches leading to *X. laevis* and Percomorphia (ray-finned fishes), WGD events have been proposed (73,74). These two branches show accelerated gene gain rates for four and two TF families, respectively. For the Gnathostomata branch no WGD has been proposed directly, but for its parent branch, the branch leading to Vertebrata, the 2R WGD events have been proposed (75,76). The only non-gnathostome vertebrate in our species set is the lamprey. The *Petromyzon marinus* genome likely caused an artifact in the ancestral reconstruction of TF family sizes because of its vertebrate-atypical small proteome size, 30% smaller than the next smallest analyzed vertebrate (*P. marinus*: 10 415 proteins, *Gallus gallus*: 15 508 proteins, no splice variants counted, from ensembl annotation). Consequently the accelerated gene gain rate on the Gnathostomata branch is likely connected to the 2R WGD events.

However, in other cases accelerated gene gain rates can not be linked to WGD events. The branches leading to *R. norvegicus* and Deuterostomia, for example, show accelerated gene gain rates for four TF families while no WGD has occurred on these branches. Other branches without WGD event show accelerated gene gain rates only for one

Table 1. Tree branches with an exceptionally high gene gain rate for one or more of the TF families and the evolutionary events that can be linked with the accelerated gene gain rate

Branch	Event	TF families
Caenorhabditis	SGD	Nuclear Receptor
Caenorhabditis elegans	SGD	Nuclear Receptor
Caenorhabditis briggsae	SGD	Homeobox, Nuclear Receptor
Chordata	SGD	bZIP
Cnidaria	SGD	p53
Deuterostomia	SGD	bHLH, bZIP, C2H2, Homeobox
Gnathostomata	WGD	bHLH, bZIP, C2H2, Homeobox, Nuclear Receptor, p53
Homo sapiens	SGD	C2H2
Loxodonta africana	SGD	p53
Mus musculus	SGD	bHLH, C2H2
Percomorpharia	WGD	Homeobox, bZIP
Rattus norvegicus	SGD	bHLH, C2H2, Nuclear Receptor, p53
Xenopus laevis	WGD	bHLH, bZIP, Homeobox, p53

For each branch the name of the node at the younger end of the branch was used as name.

or two TF families. *A priori*, WGDs would be expected to be linked to an accelerated gene gain rate in most TF families since all genes get duplicated and only families where many genes are lost afterward would show no acceleration in gene gain rates. It has previously been suggested that TF families show high retention rates after WGD events (31,77,78). Family expansion largely caused by SGD events, however, could be a sign of evolutionary pressure for innovation on the affected TF family. In such a case, not all TF families would be expected to be under this evolutionary pressure. Consequently, only few TF families would be expected to show accelerated gene gain rates on branches without WGD event. Many, but not all, branches seem to follow these patterns in our case. A low retention rate of some TF families after a WGD event can be explained by less evolutionary pressure for innovation on this TF family. For example, gene losses in some parts of the teleost fish lineage could explain the small number of TF families with accelerated gene gain rate on the Percomorphia branch in our reconstruction. On the other hand, evolutionary pressure for regulatory innovation could explain the accumulation of TF families with accelerated gene gain rate in the branches leading to, e.g. Deuterostomia, where no WGD event occurred. Nevertheless, we find that WGD events lead to accelerated TF family expansion rates for all analyzed branches with WGD in Metazoa, at least in some TF families. Additionally, we find a number of branches with increased TF family expansion rates caused by SGDs. These findings show that WGD as well as SGD both contribute to TF family expansions. To further understand the mechanism of TF family expansion, we analyzed the domain arrangements found in the TF families.

TF family size is correlated with number of DACs and unique domains

We analyzed the relationship between DRs and TF family expansion to elucidate the role of DRs for the expansion of TF families. All TF families show a positive correlation between TF family size and the number of DACs (Figure 2A). However, the strength of the correlation varies between the TF families (Table 2). The strongest correlation (0.93) can be found for the Homeobox and C2H2 ZF TF families, which are also the two largest TF families in most of the analyzed species. The increase in the number of DACs per TF family with TF family expansion could either be a by-product of the TF family evolution or a required step during TF family expansion. Given that protein domains are seen as the functional subunits of proteins it seems logical that DRs strongly influence TF function in various ways. Additional domains can also restrict the dimerization partners of dimerizing proteins and thereby modify the TF family's dimerization network (35). Creating dimerization sub-networks could facilitate functional diversification of TFs by minimizing cross-talk between different functions. An additional domain could also facilitate interaction with other molecules in the cell, i.e. signaling. The PAS domain is an example for a protein domain that can facilitate signaling in a protein (57) and can be found in the bHLH TF family (35).

Apart from additional domains, rearrangement of existing domains can also influence TF function (79). Such changes have been reported for many families, e.g. a number of plant gene families (80), many genes involved in signal transduction (81) and globins (82). In the C2H2 ZF TF family the C2H2 domain can be repeated as often as 30 times. The repetition of the DBD could in this case augment the number of possible target sequences in the DNA and thereby facilitate functional diversification. Additionally, this higher number of target sequences could allow family expansion, since previously the number of target sequences was suggested to be limiting to family size (33).

There is also a correlation between the number of unique domains and the number of genes per TF family (Figure 2 and Table 2). The implications of this correlation are quite similar to the implications of the correlation between number of DACs and number of genes. The main difference between the two analyses is that, when counting DACs, all possible arrangements of domains, i.e. repititions or changed order, are counted separately. When counting unique domains, each domain is only counted once, regardless of the number of separate arrangements it occurs in. Counting all DRs has the advantage of also considering events such as domain duplications that are common, especially in C2H2 ZFs (83,84). In practice, both measures

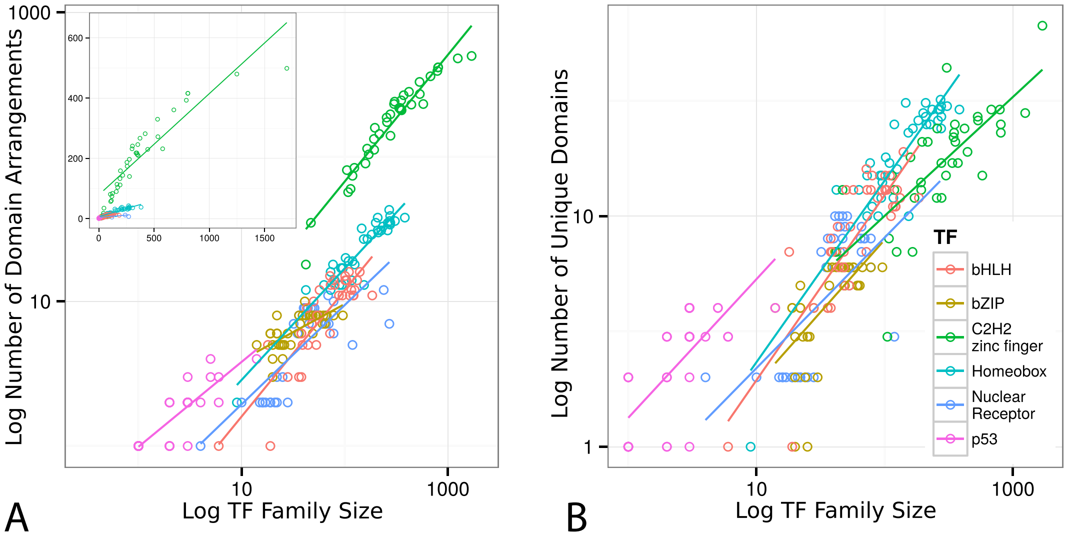


Figure 2. Relationship between domains and number of genes per TF family for all analyzed species. Linear regression lines are shown for each TF family. (**A**) Number of DACs (different domain arrangements) per TF family plotted against number of genes in the respective TF family. Full scale graph shows a log–log plot, inset shows linear axis. Each of the points represents one species. (**B**) Number of unique domains per TF family plotted against number of genes in the respective TF family. Each point represents one species.

Table 2. Correlation between TF family gene number and number of DACs in the respective TF family

TF family	Correlation to DAC number	Correlation to number of unique domains
bHLH	0.76	0.72
bZIP	0.66	0.67
C2H2 zinc finger	0.92	0.78
Homeobox	0.91	0.84
Nuclear Receptor	0.52	0.35
p53	0.79	0.56

The correlation of gene number and number of unique domains per TF family is also shown. The values given are product-moment correlation coefficients.

are meaningful, as the number of unique domains can show gain of novel functions and the number of domain arrangements can show events of major restructuring of TF proteins.

DACs are functional subunits of TF families

To determine the influence of DRs on TF function we tested the DACs of each TF family for GO term enrichment. In human, most DACs showed significant enrichment for certain GO terms, except in the C2H2 ZF family where only less than half of the DACs showed functional enrichment (Supplementary Table S1). For other species fewer DACs showed enrichment of GO terms. This result is likely caused by an incomplete annotation of TFs in species other than human. The enrichment of GO terms in the DACs shows that functions differ between the DACs of a TF family and at least some genes in each DAC share a function. The enriched GO terms of a DAC can cover a range of completely different functions (Figure 3). For example, DACs can show enrichment for GO terms as different as muscle cell differentiation and nephron tubule development. The enrichment for different GO terms shows that the genes belonging to each DAC can facilitate a wide range of functions.

The enrichment of certain GO terms in the DACs' genes could be caused by an influence of the domain arrangement on the function of proteins. An influence of domain arrangement on function would explain differences in function between the DACs. As mentioned previously, there are various ways in which changes in domain arrangements can influence TF function, e.g. by adding signaling or dimeriza-

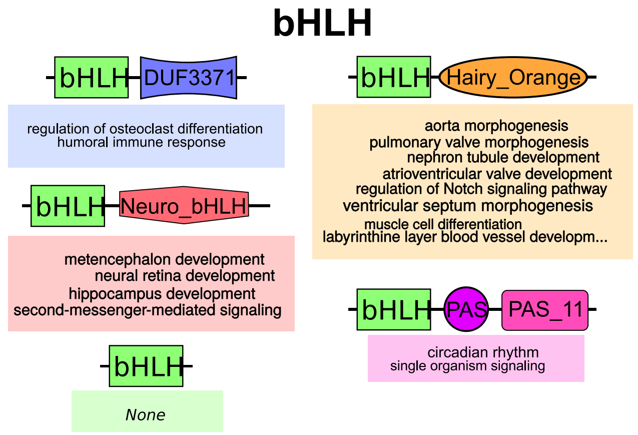

bHLH
bHLH
DUF3371
regulation of osteoclast differentiation
humoral immune response
bHLH
Hairy_Orange
aorta morphogenesis
pulmonary valve morphogenesis
nephron tubule development
atrioventricular valve development
regulation of Notch signaling pathway
ventricular septum morphogenesis
muscle cell differentiation
labyrinthine layer blood vessel developm...
bHLH
Neuro_bHLH
metencephalon development
neural retina development
hippocampus development
second-messenger-mediated signaling
bHLH
PAS
PAS_11
circadian rhythm
single organism signaling
bHLH
None

Figure 3. Wordclouds of the GO terms found to be enriched in the DACs of the bHLH TF family in human. For each DAC a pictogram of the domain arrangement is shown. Each GO term is scaled according to the *P*-value found in the enrichment test (smaller *P*-values mean bigger font size).

tion functionality to certain genes through the gain of certain domains.

Expression patterns differ between DACs

Since genes with similar expression patterns are expected to have similar functions (85–87), we analyzed the expression patterns of the TF family members. We determined if DAC members share the same expression pattern as an alternative explanation for the enrichment of GO terms in DACs found in the previous section. However, we find that TFs of a DAC rarely share the same expression pattern, i.e. many genes that have the same domain arrangement do not share the same expression pattern (see Supplementary Figure S3). Expression clusters consist of genes that all show high expression in some tissues, but low expression in the rest of the tissues. The domain arrangements of the genes found in the expression clusters differ, with several different arrangements present among them. Also, TFs with the same domain arrangement can be found in various clusters of TFs with similar expression patterns. Still, enrichment of GO terms could be found in clusters of TFs with similar expression patterns. But the GO terms enriched in clusters of TFs with similar expression patterns are different from the terms found enriched in DACs (compare Supplementary Table 4 and Figure S3). This finding suggests that DACs and expression clusters both represent functional subunits of TF families. However, these subunits are not congruent, meaning that genes with the same domain arrangement show different expression patterns that are necessary to carry out the specific regulation in multiple tissues. Additionally, members of different DACs are present in the same expression cluster. Joint expression could lead to interference between TF family members. Likely, DRs represent a mechanism via which interference can be inhibited due to changed dimerization preferences. In this way, DRs could also facilitate TF family growth.

In an additional step, we analyzed the breadth of TF expression, i.e. the number of organs a TF was found to be expressed in human (FPKM >= 1; Supplementary Figure S5). Across all TFs, most TFs were found to be expressed either in most organs or few/none of the analyzed organs. Only few of the TFs being expressed in an intermediate number of organs. Globally, this pattern has already been found in previous studies which did not differentiate TF families and DACs (26,88). However, when analyzing expression breadth of the TF families separately, our results reveal a different pattern. For the Homeobox family, for example, most genes are expressed in few tissues and only few are expressed in more than four organs. For the bZIP family, on the other hand, most genes are expressed in more than four organs. These differences in expression breadth most likely stand in relation to the TF function. Homeobox genes are often associated with developmental functions and would as such not be expected to be expressed in many adult organs. When analyzing the expression breadth of the genes in the various DACs according to the DAC's evolutionary age, the pattern visible for the whole TF family is also represented in most of the DACs (Supplementary Figure S6). There does not seem to be a relationship between DAC age and expression breadth, all patterns of expression breadth appear in all age groups. In this, our results are somewhat in contrast to previous results that proposed a more specialized expression of recently duplicated genes (89). According to our results, the C2H2 family with many recent duplications is broadly expressed. However, this might also be related to specific functions of the C2H2 family in silencing mobile elements in the genome (90,91).

CONCLUSIONS

Our results show that the expansion of TF families is often accompanied by a functional diversification that follows modular DRs. According to our findings, gene duplications offer the potential for sequence changes in one of the copies, in agreement with the established theories about gene duplications (75,92). Among the possible mutations, DRs offer the largest shift in function. By gaining a dimerization and sensing domain such as the PAS domain in bHLH, a gene copy can establish new functions such as binding signaling molecules in the cell and also act independently from the rest of the family through a new dimerization specificity. Through further gene duplications (especially WGD events), a new sub-family can be established. According to our model, WGD events per se do not add much complexity; however, functional diversification of expanded gene families after a certain time can do so.

Our study offers a solution to the riddle of how WGDs and seemingly gradual molecular changes can both help increase the complexity although WGDs alone seem to have little effect (see above). True innovation in function often requires a predating molecular change as trigger. Such a trigger can be a rearrangement of domains or the exaptation of a duplicate for a new function and both may lead to a radical shift in function. DRs and emergence as a trigger for functional shifts across a wide range of regulatory proteins have also been reported in recent studies concentrating on genomic comparisons of closely related insect species (81,93). An additional mechanism of functional diversification found in this study is change of expression patterns. These two mechanisms can help explaining the expansion of TF families by laying out how novel functions can be obtained.

Once established, such true novelties are receptive to further expansions and fine tuning which may allow for a rapid expansion of TF families and diversification of functions of family members. A possible WGD leads to a large amount of raw material which is, according to our data, in many cases rapidly utilized. However, these subsequent changes in TF protein sequence are mostly subtle, at least initially, leaving the overall architecture of regulation in order. This relationship is obvious, for example in the maintenance of interaction patterns in bZIP proteins (see above and (34)) and helps to explain why WGDs can not easily be linked to sudden organismic innovations (21,32,94). WGDs may of course still be instrumental, for example for adaptation under rapidly changing environmental conditions (29,74,95), but their adaptive value is likely not primarily related to the innovative potential of novel TFs but rather to the changes in gene expression brought by the WGDs (95). SGDs, on the other hand, can also contribute to network growth, since their duplicates also inherit their interaction preferences.

A remarkable case in point is the MADS TF family which has only five copies in human (26) and no known major expansions in any metazoan linage, but up to a hundred copies in plants (27). The MADS TF family has probably evolved by exaptation from a DNA topoisomerase (38). In plants, an array of several domains, which are mostly involved in the dimerization (or multimerization) of MADS proteins, has been acquired in a group of paralogs which became known as MIKC-type MADS proteins. These, but not any of the MIKC-free MADS proteins, then duplicated to form a dense interaction network (39). This interaction network mainly evolved from a starting point of nine to eleven interacting MIKC proteins via WGDs that left the core-interaction patterns intact (78). MIKC-type MADS proteins are key determinants of plant flower development (ABC model) and are thus instrumental for the intricacies of petal development (96). Therefore, in striking resemblance to the recruitment of domains by metazoan bHLH proteins (28), the acquisition of the IKC domains in the MADS TF family seems to have triggered a functional shift which allowed for subsequent expansion via WGDs, as was also the case in metazoan bZIP proteins (34).

In all scenarios, continuous changes in function, such as gradual shifts of sub-optimal functions as they can be observed in some enzymes (97) have not been reported for TF evolution. A possible reason may be that TF functions are more specific such that minor changes may render them non-functional and prone to rapid loss as has been hypothesized from mutational experiments on bHLH proteins (40). Modular rearrangements of domains offer a solution to this problem because readily approved subunits are recombined.

By delineating the relationships between TF family expansions, TF expression patterns and domain arrangements we make another step toward understanding the evolutionary history of Metazoa. We help explain how the TF families could expand during the evolution of Metazoa, an event that likely facilitated the evolution of more biological complexity (1–7). Our findings further our understanding of how the functional diversification of expanding TF families works in detail, namely by DRs and changes in expression pattern. This functional diversification seems necessary for family growth as it would help explain why only some genes are retained after duplication events. In detail, we find DRs and changes in expression to both contribute to functional diversification independently which we demonstrated by showing distinct GO enrichment in DACs and expression clusters. Overall, these findings shed a new light on how the evolution of more complex organisms with differing body plans and rising numbers of cell types occurred in a number of metazoan lineages.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

The authors would like to thank Amelie S. Veron and Tristan Bitard Feildel for stimulating discussions, Andreas Schüler for pilot studies and Carsten Kemena for computational support.

FUNDING

DFG [BO-2544/2-2 to EBB]; EBB ORCID [0000-0001-2345-6789]; EBB RESEARCHID [A-1563-2013]. Funding for open access charge: DFG [BO-2544/2-2].

Conflict of interest statement. None declared.

REFERENCES

1. Valentine,J.W., Collins,A.G. and Meyer,C.P. (1994) Morphological complexity increase in metazoans. *Paleobiology*, **20**, 131–142.
2. Degnan,B.M., Vervoort,M., Larroux,C. and Richards,G.S. (2009) Early evolution of metazoan transcription factors. *Curr. Opin. Genet. Dev.*, **19**, 591–599.
3. Charoensawan,V., Wilson,D. and Teichmann,S.A. (2010) Genomic repertoires of DNA-binding transcription factors across the tree of life. *Nucleic Acids Res.*, **38**, 7364–7377.
4. Miyata,T. and Suga,H. (2001) Divergence pattern of animal gene families and relationship with the Cambrian explosion. *Bioessays*, **23**, 1018–1027.
5. Levine,M. and Tjian,R. (2003) Transcription regulation and animal diversity. *Nature*, **424**, 147–151.
6. McCarthy,M.C. and Enquist,B.J. (2005) Organismal size, metabolism and the evolution of complexity in metazoans. *Evol. Ecol. Res.*, **7**, 681–696.
7. Vogel,C. and Chothia,C. (2006) Protein family expansions and biological complexity. *PLoS Comput. Biol.*, **2**, e48.
8. Wray,G.A. (2007) The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.*, **8**, 206–216.
9. Gaunt,S.J. and Paul,Y.-L. (2012) Changes in cis-regulatory elements during morphological evolution. *Biology*, **1**, 557–574.
10. Lang,D., Weiche,B., Timmerhaus,G., Richardt,S., Riaño-Pachón,D.M., Corrêa,L.G.G., Reski,R., Mueller-Roeber,B. and Rensing,S.A. (2010) Genome-wide phylogenetic comparative analysis of plant transcriptional regulation: a timeline of loss, gain, expansion, and correlation with complexity. *Genome Biol. Evol.*, **2**, 488–503.
11. Grimson,A., Srivastava,M., Fahey,B., Woodcroft,B.J., Chiang,H.R., King,N., Degnan,B.M., Rokhsar,D.S. and Bartel,D.P. (2008) Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature*, **455**, 1193–1197.
12. Technau,U. (2008) Evolutionary biology: small regulatory RNAs pitch in. *Nature*, **455**, 1184–1185.
13. Weirauch,M.T. and Hughes,T. (2011) A catalogue of eukaryotic transcription factor types, their evolutionary origin, and species distribution. In: Hughes,TR (ed). *A Handbook of Transcription Factors*. Springer, Dordrecht, Vol. **52**, pp. 25–73.
14. Tomlins,S.A., Rhodes,D.R., Perner,S., Dhanasekaran,S.M., Mehra,R., Sun,X.-W., Varambally,S., Cao,X., Tchinda,J., Kuefer,R. *et al.* (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*, **310**, 644–648.
15. Gordon,S., Akopyan,G., Garban,H. and Bonavida,B. (2005) Transcription factor YY1: structure, function, and therapeutic implications in cancer biology. *Oncogene*, **25**, 1125–1142.
16. Lynch,V.J., Leclerc,R.D., May,G. and Wagner,G.P. (2011) Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nat. Genet.*, **43**, 1154–1159.
17. Lespinet,O., Wolf,Y.I., Koonin,E.V. and Aravind,L. (2002) The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res.*, **12**, 1048–1059.
18. Babu,M.M., Luscombe,N.M., Aravind,L., Gerstein,M. and Teichmann,S.A. (2004) Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol.*, **14**, 283–291.
19. Kusserow,A., Pang,K., Sturm,C., Hrouda,M., Lentfer,J., Schmidt,H.A., Technau,U., von Haeseler,A., Hobmayer,B., Martindale,M.Q. *et al.* (2005) Unexpected complexity of the Wnt gene family in a sea anemone. *Nature*, **433**, 156–160.
20. Buitrago-Flórez,F.J., Restrepo,S. and Riaño-Pachón,D.M. (2014) Identification of transcription factor genes and their correlation with the high diversity of stramenopiles. *PLoS One*, **9**, e111841.
21. Lang,D. and Rensing,S.A. (2015) The evolution of transcriptional regulation in the viridiplantae and its correlation with morphological

complexity. In: Ruiz-Trillo,I and Nedelcu,AM (ed). *Evolutionary Transitions to Multicellular Life*. Springer, The Netherlands, pp. 301–333.

22. Albertin,C.B., Simakov,O., Mitros,T., Wang,Z.Y., Pungor,J.R., Edsinger-Gonzales,E., Brenner,S., Ragsdale,C.W. and Rokhsar,D.S. (2015) The octopus genome and the evolution of cephalopod neural and morphological novelties. *Nature*, **524**, 220–224.
23. Collén,J., Porcel,B., Carré,W., Ball,S.G., Chaparro,C., Tonon,T., Barbeyron,T., Michel,G., Noel,B., Valentin,K. *et al.* (2013) Genome structure and metabolic features in the red seaweed Chondrus crispus shed light on evolution of the Archaeplastida. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 5247–5252.
24. Jovelin,R. (2009) Rapid sequence evolution of transcription factors controlling neuron differentiation in caenorhabditis. *Mol. Biol. Evol.*, **26**, 2373–2386.
25. de Mendoza,A., Sebé-Pedrós,A., Šestak,M.S., Matejčić,M., Torruella,G., Domazet-Lošo,T. and Ruiz-Trillo,I. (2013) Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, E4858–E4866.
26. Vaquerizas,J.M., Kummerfeld,S.K., Teichmann,S.A. and Luscombe,N.M. (2009) A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.*, **10**, 252–263.
27. Becker,A. and Theißen,G. (2003) The major clades of MADS-box genes and their role in the development and evolution of flowering plants. *Mol. Phylogenet. Evol.*, **29**, 464–489.
28. Amoutzias,G.D., Robertson,D.L., Oliver,S.G. and Bornberg-Bauer,E. (2004) Convergent evolution of gene networks by single-gene duplications in higher eukaryotes. *EMBO Rep.*, **5**, 274–279.
29. Van de Peer,Y., Maere,S. and Meyer,A. (2009) The evolutionary significance of ancient genome duplications. *Nat. Rev. Genet.*, **10**, 725–732.
30. Smet,R.D., Adams,K.L., Vandepoele,K., Montagu,M.C.E.V., Maere,S. and de Peer,Y.V. (2013) Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 2898–2903.
31. Rensing,S.A. (2014) Gene duplication as a driver of plant morphogenetic evolution. *Curr. Opin. Plant Biol.*, **17**, 43–48.
32. Crow,K.D. and Wagner,G.P. (2006) What is the role of genome duplication in the evolution of complexity and diversity? *Mol. Biol. Evol.*, **23**, 887–892.
33. Itzkovitz,S., Tlusty,T. and Alon,U. (2006) Coding limits on the number of transcription factors. *BMC Genomics*, **7**, 239.
34. Amoutzias,G.D., Veron,A.S., Weiner,J., Robinson-Rechavi,M., Bornberg-Bauer,E., Oliver,S.G. and Robertson,D.L. (2007) One Billion Years of bZIP Transcription Factor Evolution: Conservation and Change in Dimerization and DNA-Binding Site Specificity. *Molecular Biology and Evolution*, **24**, 827–835.
35. Amoutzias,G.D., Robertson,D.L., Van de Peer,Y. and Oliver,S.G. (2008) Choose your partners: dimerization in eukaryotic transcription factors. *Trends Biochem. Sci.*, **33**, 220–229.
36. Roberts,D., Keeling,R., Tracka,M., van der Walle,C.F., Uddin,S., Warwicker,J. and Curtis,R. (2014) The role of electrostatics in protein–protein interactions of a monoclonal antibody. *Mol. Pharm.*, **11**, 2475–2489.
37. Roberts,D., Keeling,R., Tracka,M., van der Walle,C.F., Uddin,S., Warwicker,J. and Curtis,R. (2015) Specific ion and buffer effects on protein–protein interactions of a monoclonal antibody. *Mol. Pharm.*, **12**, 179–193.
38. Gramzow,L., Ritz,M.S. and Theißen,G. (2010) On the origin of MADS-domain transcription factors. *Trends Genet.*, **26**, 149–153.
39. Kaufmann,K., Melzer,R. and Theißen,G. (2005) MIKC-type MADS-domain proteins: structural modularity, protein interactions and network evolution in land plants. *Gene*, **347**, 183–198.
40. Maerkl,S.J. and Quake,S.R. (2009) Experimental determination of the evolvability of a transcription factor. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 18650–18655.
41. Erwin,D.H., Laflamme,M., Tweedt,S.M., Sperling,E.A., Pisani,D. and Peterson,K.J. (2011) The cambrian conundrum: early divergence and later ecological success in the early history of animals. *Science*, **334**, 1091–1097.
42. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Billis,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fitzgerald,S. *et al.* (2014) Ensembl 2014. *Nucleic Acids Res.*, **42**, D749–D755.
43. Kersey,P.J., Allen,J.E., Christensen,M., Davis,P., Falin,L.J., Grabmueller,C., Hughes,D. S.T., Humphrey,J., Kerhornou,A., Khobova,J. *et al.* (2014) Ensembl Genomes 2013: scaling up access to genome-wide data. *Nucleic Acids Res.*, **42**, D546–D552.
44. Punta,M., Coggill,P.C., Eberhardt,R.Y., Mistry,J., Tate,J., Boursnell,C., Pang,N., Forslund,K., Ceric,G., Clements,J. *et al.* (2011) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.
45. Csűös,M. (2010) Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics*, **26**, 1910–1912.
46. R Core Team and others. (2012) R: a language and environment for statistical computing. http://cran.case.edu/web/packages/dplR/vignettes/timeseries-dplR.pdf.
47. Wickham,H. (2009) *ggplot2: Elegant Graphics for Data Analysis*, Springer, NY.
48. Hunter,J.D. (2007) Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.*, **9**, 90–95.
49. Alexa,A. and Rahnenführer,J. (2010) topGO: enrichment analysis for gene ontology. *R package version 2.12*. Available from: http://www.bioconductor.org/packages/2.12/bioc/html/topGO.html.
50. Brawand,D., Soumillon,M., Necsulea,A., Julien,P., Csárdi,G., Harrigan,P., Weier,M., Liechti,A., Aximu-Petri,A., Kircher,M. *et al.* (2011) The evolution of gene expression levels in mammalian organs. *Nature*, **478**, 343–348.
51. Petryszak,R., Burdett,T., Fiorelli,B., Fonseca,N.A., Gonzalez-Porta,M., Hastings,E., Huber,W., Jupp,S., Keays,M., Kryvych,N. *et al.* (2014) Expression Atlas update—a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Res.*, **42**, D926–D932.
52. Wagner,G.P., Kin,K. and Lynch,V.J. (2012) Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.*, **131**, 281–285.
53. Wild,F. (2014) lsa: latent semantic analysis. http://CRAN.R-project.org/package=lsa.
54. Jones,E., Oliphant,T., Peterson,P. *et al.* (2001) *SciPy: Open source scientific tools for Python*. http://www.scipy.org/ (28 January 2016, date last accessed).
55. Morgenstern,B. and Atchley,W.R. (1999) Evolution of bHLH transcription factors: modular evolution by domain shuffling? *Mol. Biol. Evol.*, **16**, 1654–1663.
56. McIntosh,B.E., Hogenesch,J.B. and Bradfield,C.A. (2010) Mammalian Per-Arnt-Sim proteins in environmental adaptation. *Annu. Rev. Physiol.*, **72**, 625–645.
57. Mei,Q. and Dvornyk,V. (2014) Evolution of PAS domains and PAS-containing genes in eukaryotes. *Chromosoma*, **123**, 385–405.
58. Bornberg-Bauer,E., Rivals,E. and Vingron,M. (1998) Computational approaches to identify leucine zippers. *Nucleic Acids Res.*, **26**, 2740–2746.
59. Karin,M., Liu,Z.-g. and Zandi,E. (1997) AP-1 function and regulation. *Curr. Opin. Cell Biol.*, **9**, 240–246.
60. Shaulian,E. and Karin,M. (2001) AP-1 in cell proliferation and survival. *Oncogene*, **20**, 2390–2400.
61. Bürglin,T.R. (2011) Homeodomain subtypes and functional diversity. In: Hughes,TR (ed). *A Handbook of Transcription Factors*. Springer, The Netherlands, pp. 95–122.
62. Gehring,W.J., Affolter,M. and Burglin,T. (1994) Homeodomain proteins. *Annu. Rev. Biochem.*, **63**, 487–526.
63. Robinson-Rechavi,M., Garcia,H.E. and Laudet,V. (2003) The nuclear receptor superfamily. *J. Cell Sci.*, **116**, 585–586.
64. Pardee,K., Necakov,A.S. and Krause,H. (2011) Nuclear receptors: small molecule sensors that coordinate growth, metabolism and reproduction. In: Hughes,TR (ed). *A Handbook of Transcription Factors*, Springer, Dordrecht, pp. 123–153.
65. Glass,C.K. (1994) Differential recognition of target genes by nuclear receptor monomers, dimers, and heterodimers. *Endocr. Rev.*, **15**, 391–407.
66. Görner,W., Durchschlag,E., Martinez-Pastor,M.T., Estruch,F., Ammerer,G., Hamilton,B., Ruis,H. and Schüller,C. (1998) Nuclear localization of the C2H2 zinc finger protein Msn2p is regulated by stress and protein kinase A activity. *Genes Dev.*, **12**, 586–597.

67. Nowick,K., Fields,C., Gernat,T., Caetano-Anolles,D., Kholina,N. and Stubbs,L. (2011) Gain, loss and divergence in primate zinc-finger genes: a rich resource for evolution of gene regulatory differences between species. *PLoS One*, **6**, e21553.
68. Nowick,K., Carneiro,M. and Faria,R. (2013) A prominent role of KRAB-ZNF transcription factors in mammalian speciation? *Trends Genet.*, **29**, 130–139.
69. Ho,W.C., Fitzgerald,M.X. and Marmorstein,R. (2006) Structure of the p53 core domain dimer bound to DNA. *J. Biol. Chem.*, **281**, 20494–20502.
70. Levine,A.J. and Oren,M. (2009) The first 30 years of p53: growing ever more complex. *Nat. Rev. Cancer*, **9**, 749–758.
71. Reece-Hoyes,J.S., Deplancke,B., Shingles,J., Grove,C.A., Hope,I.A. and Walhout,A.J. (2005) A compendium of Caenorhabditis elegans regulatory transcription factors: a resource for mapping transcription regulatory networks. *Genome Biol.*, **6**, R110.
72. Abegglen,L.M., Caulin,A.F. and Chan,A. (2015) Potential mechanisms for cancer resistance in elephants and comparative cellular response to dna damage in humans. *JAMA*, **314**, 1850–1860.
73. Uno,Y., Nishida,C., Takagi,C., Ueno,N. and Matsuda,Y. (2013) Homoeologous chromosomes of *Xenopus laevis* are highly conserved after whole-genome duplication. *Heredity*, **111**, 430–436.
74. Meyer,A. and Van de Peer,Y. (2005) From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). *Bioessays*, **27**, 937–945.
75. Ohno,S. (1970) *Evolution by gene duplication*. George Alien & Unwin Ltd/Springer-Verlag, London/Berlin, Heidelberg and NY.
76. Sidow,A. (1996) Gen(om)e duplications in the evolution of early vertebrates. *Curr. Opin. Genet. Dev.*, **6**, 715–722.
77. Maere,S., Bodt,S.D., Raes,J., Casneuf,T., Montagu,M.V., Kuiper,M. and de Peer,Y.V. (2005) Modeling gene and genome duplications in eukaryotes. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 5454–5459.
78. Veron,A.S., Kaufmann,K. and Bornberg-Bauer,E. (2007) Evidence of interaction network evolution by whole-genome duplications: a case study in MADS-Box proteins. *Mol. Biol. Evol.*, **24**, 670–678.
79. Björklund,A.K., Ekman,D., Light,S., Frey-Skött,J. and Elofsson,A. (2005) Domain rearrangements in protein evolution. *J. Mol. Biol.*, **353**, 911–923.
80. Kersting,A.R., Mizrachi,E., Bornberg-Bauer,E. and Myburg,A.A. (2015) Protein domain evolution is associated with reproductive diversification and adaptive radiation in the genus Eucalyptus. *New Phytol.*, **206**, 1328–1336.
81. Moore,A.D., Grath,S., Schüler,A., Huylmans,A.K. and Bornberg-Bauer,E. (2013) Quantification and functional analysis of modular protein evolution in a dense phylogenetic tree. *Biochim. Biophys. Acta*, **1834**, 898–907.
82. Projecto-Garcia,J., Jollivet,D., Mary,J., Lallier,F.H., Schaeffer,S.W. and Hourdez,S. (2015) Selective forces acting during multi-domain protein evolution: the case of multi-domain globins. *Springerplus*, **4**, 354.
83. Iuchi,S. (2001) Three classes of C2H2 zinc finger proteins. *Cell. Mol. Life Sci.*, **58**, 625–635.
84. Stubbs,L., Sun,Y. and Caetano-Anolles,D. (2011) Function and evolution of C2H2 zinc finger arrays. In: Hughes,TR (ed). *A Handbook of Transcription Factors*. Springer, Dordrecht, pp. 75–94.
85. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 14863–14868.
86. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
87. Stuart,J.M., Segal,E., Koller,D. and Kim,S.K. (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249–255.
88. Freilich,S., Massingham,T., Bhattacharyya,S., Ponsting,H., Lyons,P.A., Freeman,T.C. and Thornton,J.M. (2005) Relationship between the tissue-specificity of mouse gene expression and the evolutionary origin and function of the proteins. *Genome Biol.*, **6**, R56.
89. Freilich,S., Massingham,T., Blanc,E., Goldovsky,L. and Thornton,J.M. (2006) Relating tissue specialization to the differentiation of expression of singleton and duplicate mouse proteins. *Genome Biol.*, **7**, R89.
90. Thomas,J.H. and Schneider,S. (2011) Coevolution of retroelements and tandem zinc finger genes. *Genome Res.*, **21**, 1800–1812.
91. Jacobs,F. M.J., Greenberg,D., Nguyen,N., Haeussler,M., Ewing,A.D., Katzman,S., Paten,B., Salama,S.R. and Haussler,D. (2014) An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature*, **516**, 242–245.
92. Zhang,J. (2003) Evolution by gene duplication: an update. *Trends Ecol. Evol.*, **18**, 292–298.
93. Moore,A.D. and Bornberg-Bauer,E. (2012) The dynamics and evolutionary potential of domain loss and emergence. *Mol. Biol. Evol.*, **29**, 787–796.
94. Donoghue,P. C.J. and Purnell,M.A. (2005) Genome duplication, extinction and vertebrate evolution. *Trends Ecol. Evol.*, **20**, 312–319.
95. Fawcett,J.A., Maere,S. and de Peer,Y.V. (2009) Plants with double genomes might have had a better chance to survive the Cretaceous–Tertiary extinction event. *Proc. Natl. Acad. Sci.*, **106**, 5737–5742.
96. Honma,T. and Goto,K. (2001) Complexes of MADS-box proteins are sufficient to convert leaves into floral organs. *Nature*, **409**, 525–529.
97. Voordeckers,K., Brown,C.A., Vanneste,K., van der Zande,E., Voet,A., Maere,S. and Verstrepen,K.J. (2012) Reconstruction of ancestral metabolic enzymes reveals molecular mechanisms underlying evolutionary innovation through gene duplication. *PLoS Biol.*, **10**, e1001446.