Bioinformatic Studies on Selection in Protein-Coding- and RNA-Specifying Genes

A Dissertation

Presented to

the Faculty of the Department of Biology and Biochemistry

University of Houston

In Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By

Nicholas Price

May 2014

Bioinformatic Studies on Selection in Protein-Coding- and RNA-Specifying Genes

Nicholas Price

APPROVED:

Dr. Dan Graur, Chair

Dr. Ricardo Azevedo

Dr. Tim Cooper

Dr. Kresimir Josic

Dr. Michael Kohn

Dr. Dan Wells, Dean, College of Natural Sciences and Mathematics

Acknowledgments

I thank Dr. Ricardo Azevedo, Dr. Niv Sabath and Dr. Reed Cartwright for sharing their enthusiasm in science and knowledge in molecular evolution. Collaborating with them further increased my understanding of some basic principles in molecular evolution and statistical analysis.

I want to especially thank Dr. Kiyoshi Ezawa for the great scientific discussions we had over the years and recent collaborations that really enhanced my critical thinking. Dr. Kiyoshi Ezawa taught me that as a scientist you should be critical of scientific work but always keep an open mind.

Finally, I owe a great thanks to my advisor and friend Dr. Dan Graur. Dr. Dan Graur expects his students to be very independent. At times, this caused me to be frustrated blaming myself and him, but towards the end of my PhD I realized that thinking independently, facing problems along the way, and being patient and persistent are all necessary ingredients for a successful scientist. By observing Dr. Dan Graur as a scientist, I learned that when writing a paper every statement made should be supported by the data; nothing more or nothing less. Furthermore, I learned that when testing a hypothesis I should be very confident that the statistical procedures used are the appropriate ones. Besides being a great advisor, Dr. Dan Graur has been a friend. I really

enjoyed talking to him about life, politics and art. Although my appreciation of art has grown immensely since I started my PhD I still have a long way to go.

Bioinformatic Studies on Selection in Protein-Coding- and RNA-Specifying Genes

An Abstract of a Dissertation

Presented to

the Faculty of the Department of Biology and Biochemistry

University of Houston

In Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By

Nicholas Price

May 2014

Abstract

During my dissertation I used a combination of comparative and phylogenetic approaches to test for selection on genic regions. Specifically, I tested for mutational and environmental robustness in *Drosophila* pre-miRNAs, as well as selection on synonymomous and nonsynonymous sites in mammalian protein coding genes. Contrary to previous claims of selection for mutational robustness in mammalian pre-miRNAs, I demonstrate that mutational robustness evolved neutrally in *Drosophila* pre-miRNas. Furthermore, I show that mutational robustness did not evolve as a byproduct of selection for environmental robustness. In Chapter 3, I identify orthologous processed pseudogenes and use them to test for selection on synonymous sites. By estimating the rate of substitution at synonymous sites in genes (d_S), and corresponding sites in pseudogenes (d_ψ), I demonstrate that only about ~8% of synonymous changes could possibly be under selection. This is in stark contrast to the ~30% previously claimed. However, I show that both deviations from neutrality can be caused by an increase in divergence between the sequences examined, in combination, with a difference in nucleotide composition between genes and pseudogenes. In the last part of my dissertation, I estimate selection on nonsynonymous sites in mammals (d_N/d_S), and examine the previously claimed positive correlation between generation time and selection. To ensure the correct estimation of selection, I examine the effects of sequencing errors and alignment quality. After accounting for phylogenetic independence, I find no correlation between generation time and selection. Furthermore, I find a significant decrease in the efficiency of selection in monkeys, after the simian and prosimian split.

Contents

Chapter One: General introduction

Mutations occurring in functional DNA sequences may or may not alter the phenotype of an organism. In the event that they do, they may or may not affect the fitness of the organism that carries the mutation. Mutations can be deleterious, neutral or advantageous. In mammals, in which genomes contain vast amounts of junk DNA, most new mutations occur in nonfunctional regions of the genome and are, hence, neutral, i.e., are as fit as the fittest allele in the population. Most non-neutral mutations occurring within a population reduce the fitness of their carriers, i.e., are deleterious. Deleterious mutations are selected against and eventually removed from the population. This type of selection is called negative or purifying. Very rarely, a mutation is advantageous and increases the fitness of its carriers. Such a mutation will increase in frequency, that is, it will be subjected to positive or advantageous selection (Graur 2014).

The fate of neutral alleles in a population is determined solely by random genetic drift, while the fate of deleterious and advantageous alleles could be affected by both random genetic drift and selection.

To assesses the past effects of random genetic drift on a population, one can measure the neutral genetic diversity between a sample of individuals and determine the effective population size (N_e). The effective population size, is the size of a so called “ideal” Wright-Fisher population (Wright 1931) that would experience the effects of genetic drift to the same degree as the population under study.

In 1962, Kimura derived the probability of fixation of a new allele in a diploid population given the effective population size N_e, the census population size N, and the selection coefficient, s, assuming codominance (Kimura 1962).

(1.1) $$P = \frac{1 - e^{-(2N_e s/N)}}{1 - e^{-(4N_e s)}}$$

The selection coefficient is defined as the relative fitness of a phenotype relative to another phenotype. As shown in Figure 1.1a the probability of fixation of an allele with a selective disadvantage of -0.001 decreases as the effective population size approaches the census population size. On the other hand, the probability of fixation of an advantageous allele with ($s = 0.01$) increases Figure 1.1b.

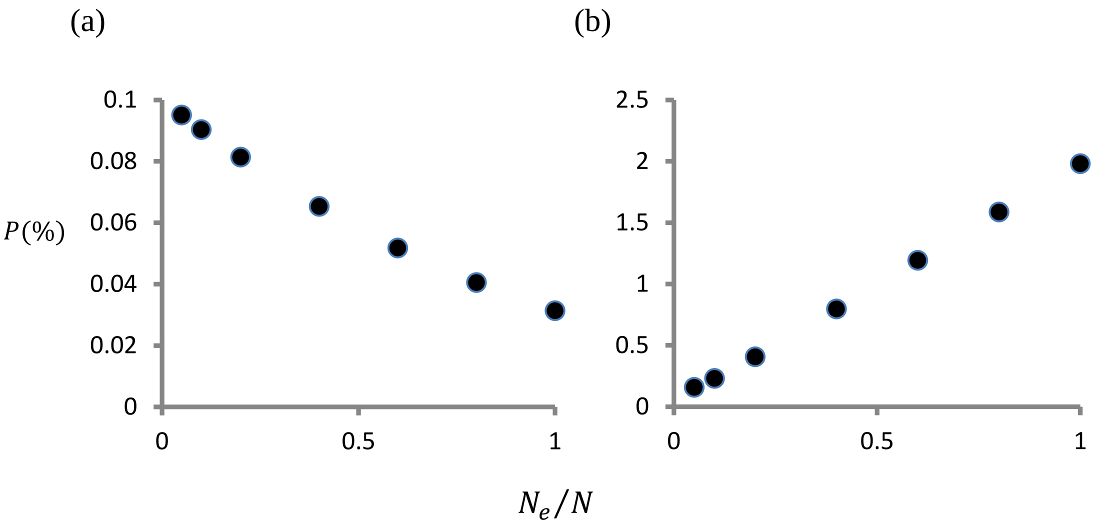

Figure 1.1: The probability of fixation of a deleterious (left) and advantageous allele (right) as effective population size (N_e) approaches the census population size (N). N=500 individuals

These trends show that the efficiency of selection increases as random genetic drift decreases. According to theory, in a diploid population, selection for an advantageous mutation or selection against a deleterious mutation is efficient only when $2N_es \gg 1$ (Kimura 1968a). If this criterion is not met, then the fate of a mutation is largely determined by random genetic drift.

In this dissertation I use both comparative and phylogenetic methods to test for selection on functional sequences within eukaryotes. Given that the efficiency of selection is dependent on the effective population size, in Chapter 2, I examine whether there is selection for mutational or environmental robustness in Drosophila miRNAs. In Chapter 3, I identify processed pseudogenes and use them as a neutral model to test for selection on synonymous sites (Chapter 4). Finally, in Chapter 5, using generation time as a proxy to effective population size, I examine whether there is a relation between the efficiency of selection on nonsynonymous sites and generation time.

As mentioned above, in Chapter 2, I estimate selection for mutational and environmental robustness on precursor miRNAs (pre-miRNAs). Robustness, is the extent to which a genotype can produce the same phenotype in the face of perturbations (de Visser et al. 2003). These perturbations can be mutations or environmental factors such temperature, salinity, or pH. Robustness has been observed at all levels of genetic organization, starting from individual genes (Borenstein and Ruppin 2006; Lind et al. 2010; Hietpas et al. 2011) to whole organisms (von Dassow et al. 2000; Baba et al. 2006; Ritter et al. 2013). An example of robustness is the finding that functional genes while under selection are not always essential. For example, in *E. coli*, 90% of single gene knock outs did not produce any lethal phenotypes (Baba et al. 2006). In ribosomal protein genes, 95% of directed mutations were only weakly deleterious (Lind et al. 2010).

The two theories proposed to explain the existence of robustness are: (1) robustness offers a fitness advantage and therefore is directly selected for (de Visser et al. 2003), and (2) robustness is a correlated byproduct of selection for another property such as function (Gibson and Wagner 2000; Wagner 2005).

Recently, robustness has been observed in precursor miRNAs (Wagner and Stadler 1999; Bonnet et al. 2004; Sanjuan et al. 2007; Shu et al. 2008; Szollosi and Derenyi 2009; Churkin et al. 2010). These are sequences that fold in a stem-loop structure and contain the mature miRNA that is latter excised by the Drosha complex (Figure 1.2) (Bartel 2004). Mature miRNAs are known to regulate the expression of genes (Bartel 2004).

Robustness in pre-miRNAs has been suggested to evolve due to direct selection for resilience against mutations (Borenstein and Ruppin 2006; Sanjuan et al. 2007), or, selection for resistance against thermal fluctuations (Ancel and Fontana 2000; Szollosi and Derenyi 2009).

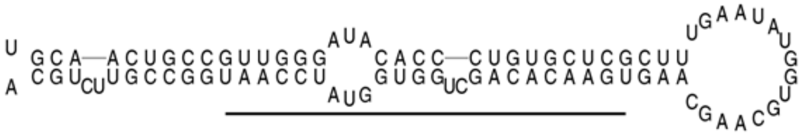

Figure 1.2: Secondary structure of mir-317 in *Drosophila Pseudoobscura*. The pre-miRNA folds into a stem-loop hairpin structure. The horizontal line marks the location of the mature miRNA.

Mutational robustness of a pre-miRNA can be predicted *in silico* by randomly mutating a pre-miRNA sequence and using an algorithm to predict the minimum free energy (MFE) structure. A mutationally robust pre-miRNA is one whose initial MFE structure is not heavily distorted after mutation. Thermal or environmental robustness can also be predicted *in silico* by studying the thermodynamic ensemble of a pre-miRNA. A thermodynamic ensemble represents the structures a pre-miRNA could fold at a given temperature. A pre-miRNA with low structural variation within its ensemble is defined as thermally robust.

To test for selection for mutational or environmental robustness previous studies have compared the robustness of real pre-miRNAs, to the robustness of random sequences folding in the same stem-loop structure (Borenstein and Ruppin 2006; Szollosi and Derenyi 2009). In the current study, I use a different approach, and study the evolution of pre-miRNA robustness over the *Drosophila* phylogeny. Specifically, I test for direct selection on mutational and environmental robustness.

In Chapter 3, I survey completely sequenced mammalian genomes, and identify processed psudogenes in mammals. These pseudogenes are later used to infer orthologous relationships. Processed pseudogenes are regions that are assumed to be non-functional and to evolve neutrally (Graur 2014), therefore, in Chapter 4, I use orthologous processed pseduogenes as a neutral model to test for selection on synonymous sites.

Mutations in protein coding genes can be classified into nonsynonymous, synonymous and nonsense. Nonsynonymous mutations lead to changes at the amino acid level, while synonymous mutations do not change the same amino acid. Although mutations at synonymous sites do not affect the structure or function of a protein, recent studies have identified that synonymous sites are involved in functions, such as : (1) harboring signals that aid in the accurate splicing out of introns (Fairbrother et al. 2004; Carlini and Genut 2006; Dewey et al. 2006; Parmley et al. 2006; Caceres and Hurst 2013), (2) stabilizing mRNA structures to avoid their degradation (Shen et al. 1999; Buratti and Baralle 2004; Chamary and Hurst 2005; Shabalina et al. 2006), (3) serving as binding sites for miRNAs and transcription factors (Hurst 2006; Gu et al. 2012; Stergachis et al. 2013) and (4) being part of codons that are more efficient and accurate during translation (Ikemura 1985; Akashi and Eyre-Walker 1998; Duret 2002; Wright et al. 2004; Stoletzki and Eyre-Walker 2007; Drummond and Wilke 2008).

Assuming that most mutations are deleterious and that synonymous sites are functional, selection on synonymous sites can still be effectively neutral if the product of the effective population size and selective disadvantage is not sufficiently large. In mammals, whose effective population sizes are low, evidence of selection against synonymous mutations has been inconsistent (Chamary et al. 2006). Specifically, the proportion of synonymous sites estimated to be under selection ranges from 1-39% (Ophir et al. 1999; Bustamante et al. 2002; Hellmann et al. 2003; Parmley et al. 2006; Eory et al. 2010; Caceres and Hurst 2013)

In Chapter 4, I use a large sample of orthologous processed pseudogenes to estimate selection on synonymous sites in primates and rodents.

Although selection on synonymous sites has been highly controversial, selection on nonsynonymous sites is clearly evident. Under the assumptions that synonymous sites are evolve neutrally, and that mutations at nonsynonymous sites are mostly deleterious, the rate of nonsynonymous substitution (d_N) should be smaller than the rate of synonymous substitution (d_S). Indeed, when using human-chimp orthologous genes, d_N, is estimated to be ten times smaller than d_S (The Chimpanzee Sequencing and Analysis Consortium 2005). The d_N/d_S ratio, otherwise abbreviated as ω, is frequenctly used to estimate selection on nonsynonymous sites, using synonymous sites as a neutral model of evolution (Li 1997).

Because the efficiency of selection against deleterious mutations increases with effective population size, it is expected that ω should decrease. This was first hypothesized by Ohta and Kimura in 1971 and then was formally presented as the nearly neutral theory of molecular evolution (Ohta 1973). Evidence of the above theory was first shown by Ohta in 1972 using generation time as a proxy to effective population size. Ohta, depicted that species with short generation times or large effective population sizes exhibit a lower rate of amino acid substitutions per nucleotide substitutions than species with longer generation times or low effective population sizes. Later studies confirmed the positive correlation between ω and generation time by comparing orthologous genes between mammals and insects. These however, were limited to a small number of genes and comparisons were mainly made between species with different life history traits. Differences in ω between species with different life history traits could result from differences in selection (s) and not effective population sizes (N_e).

With the recent sequencing of a large number of genomes, a few studies have shown that there is a positive linear relation between ω and generation time within mammals (Figure 1.4) (Nikolaev et al. 2007; Popadin et al. 2013)

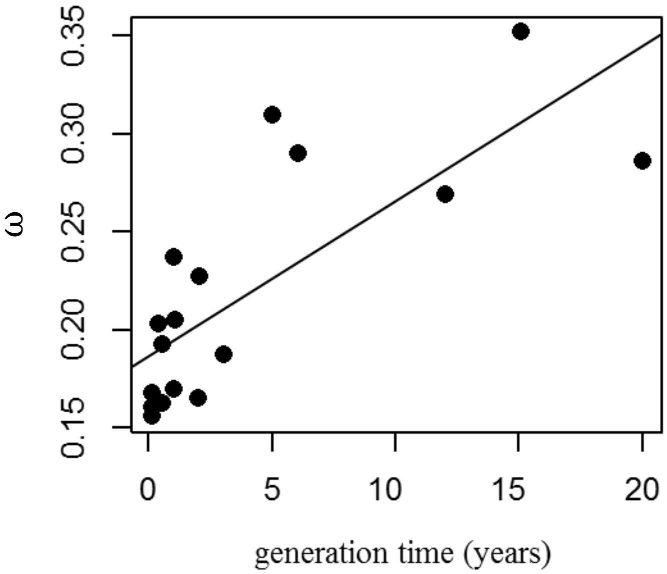

Figure 1.4: A plot of the *ω* ratio against generation time (years) in 17 mammals. Data were taken from the study by Nikolaev et al. (2007).

In Chapter 5, I estimate selection on nonsynonymous sites in 13 high coverage mammals. In estimating selection, I consider alignment quality and sequence quality both of which can have significant effects when estimating *ω* (Schneider et al. 2009).

Chapter Two: Neutral evolution of robustness in *Drosophila* microRNA precursors

Abstract

Mutational robustness describes the extent to which a phenotype remains unchanged in the face of mutations. Theory predicts that the strength of direct selection for mutational robustness is at most the magnitude of the rate of deleterious mutation. As far as nucleic-acid sequences are concerned, only long sequences in organisms with high deleterious mutation rates and large population sizes are expected to evolve mutational robustness. Surprisingly, recent studies have concluded that molecules that meet none of these conditions—the microRNA precursors (pre-miRNAs) of multicellular eukaryotes—show signs of selection for mutational and/or environmental robustness. To resolve the apparent disagreement between theory and these studies, we have reconstructed the evolutionary history of *Drosophila* pre-miRNAs and compared the robustness of each sequence to that of its reconstructed ancestor. In addition, we "replayed the tape" of pre-miRNA evolution via simulation under different evolutionary assumptions and compared these alternative histories with the actual one. We found that *Drosophila* pre-miRNAs have evolved under strong purifying selection against changes in secondary structure. Contrary to earlier claims, there is no evidence that pre-miRNAs have been shaped by direct selection for any kind of robustness.

Introduction

Robustness or canalization is the extent to which a genotype can produce the same phenotype in the face of perturbations (Gibson and Wagner 2000; Meiklejohn and Hartl 2002; de Visser et al. 2003; Flatt 2005; Wagner 2005). These perturbations can be genetic, such as mutation, recombination, and horizontal gene transfer, or environmental, such as fluctuations in temperature, food availability, or salinity. Mutational robustness is thought to be a fundamental property of biological systems, from individual molecules to gene regulatory networks (de Visser et al. 2003; Stelling et al. 2004; Kitano 2004a, 2004b; Wagner 2005) For example, Guo et al. (2004) found that 74% of nucleotide substitutions preserved at least some of the function of human enzyme 3-methyladenine DNA glycosylase (3MDG). High tolerance against mutations has been observed in many other proteins (Miller 1979; Datta and Jinks-Robertson 1995; Reddy et al. 1998; Bloom et al. 2005). In addition, conserved elements of secondary structure from the genomes of RNA viruses were found to be significantly more resistant to mutations than nonconserved elements (Wagner and Stadler 1999).

How did this high mutational robustness evolve? One possibility is that it resulted from direct selection for high mutational robustness (de Visser et al. 2003). The strength of selection for mutational robustness is at most the magnitude of the deleterious mutation rate (*U*) (Kimura 1967; Proulx and Phillips 2005). For a single RNA or protein molecule, the deleterious mutation rate is given by $U = \mu L P_{del}$, where μ is the mutation rate per site, per generation, L is the length of the sequence, and P_{del} is the probability that a mutation is deleterious. (Note that 1–P_{del} is a measure of mutational robustness.) For example, for the human enzyme 3MDG, we have L = 894 nucleotides (nt) and P_{del} = 26% (Guo et al. 2004). Assuming that μ = 2.5×10^{-8} per base pair per generation (Nachman and Crowell 2000), we estimate that $U \approx 5.8\times10^{-6}$. Thus, selection for mutational robustness is expected to be weak in human 3MDG, as well as in the vast majority of individual gene products.

The main factor determining the extent to which mutational robustness will respond to direct selection is the effective population size (Kimura 1968a; Wagner et al. 1997; van Nimwegen et al. 1999; Wilke et al. 2001; Azevedo et al. 2006; Forster et al. 2006). For example, a diploid population is expected to respond provided that it obeys the condition $2N_eU \gg 1$ (Wright 1931; Kimura 1968b; Li 1978). Therefore, according to theory, mutational robustness should only evolve under direct selection in taxa with high N_eU such as certain RNA viruses, prokaryotes and unicellular eukaryotes (Drake et al. 1998; Lynch and Conery 2003). In agreement with this prediction, experimental evidence for evolution of mutational robustness under direct selection has only been observed in an RNA virus (Montville et al. 2005; Sanjuan et al. 2007). In contrast, the mutational robustness of individual protein or RNA molecules is expected to be effectively neutral in most multicellular eukaryotes (Lynch and Conery 2003), suggesting that direct selection is an unlikely explanation for the findings of high mutational robustness (Wagner et al. 1997; van Nimwegen et al. 1999; Wilke et al. 2001; Azevedo et al. 2006; Forster et al. 2006). An alternative explanation, known as congruent selection, is that mutational robustness evolves as a by-product of selection for another form of robustness (Ancel and

Fontana 2000; Meiklejohn and Hartl 2002; de Visser et al. 2003; Wagner 2005), such as thermodynamic stability (Ancel and Fontana 2000), robustness to recombination (Azevedo et al. 2006; Gardner and Kalinka 2006; Misevic et al. 2006; Szollosi and Derenyi 2008), or robustness to transcriptional or translational errors (Ninio 1991; Wilke and Drummond 2006). For example, RNA molecules alternate rapidly among several different low-energy secondary structures. Some molecules are more thermodynamically stable than others at a constant temperature. Using computer simulations, Ancel and Fontana (2000) showed that the thermodynamic stability of an RNA molecule is positively correlated with its robustness to mutation, such that selection for the ability to produce a given structure at constant temperature caused both thermodynamic stability and mutational robustness to increase. Recently, Montville et al. (2005) demonstrated that mutational robustness evolved congruently in strains of an RNA virus selected for high and low levels of co-infection.

In recent years, microRNA precursors (pre-miRNAs) have emerged as a model system for the study of the evolution of robustness (Bonnet et al. 2004; Borenstein and Ruppin 2006; Shu et al. 2007; Shu et al. 2008; Szollosi and Derenyi 2009). The pre-miRNAs of multicellular eukaryotes are not expected to respond to direct selection for mutational robustness because they are small molecules in organisms with small populations: if we assume L = 100 and P_{del} = 100% (the maximum possible value), we expect that $2N_eU \approx 2N_e\mu$ to range from 0.054 to 0.74 for the pre-miRNAs of human, mouse, *Drosophila melanogaster,* and *Caenorhabditis elegans* (Lynch and Conery 2003).

Surprisingly, Borenstein and Ruppin (2006) reported evidence for direct selection for mutational robustness in pre-miRNAs from, among others, the above-listed species. They found that these RNAs have a higher mutational robustness than random sequences with the same secondary structure, even after controlling for the intrinsic robustness arising from the pre-miRNA hairpin structure and correcting for nucleotide composition bias. Furthermore, real and random pre-miRNAs did not differ significantly in their thermodynamic stability, which led them to conclude that the high mutational robustness was caused by direct, rather than congruent, selection (Borenstein and Ruppin 2006). If correct, these results would imply that, either current population genetics theory is wrong, or that we have grossly underestimated the effective population sizes and/or the deleterious mutation rates in multicellular eukaryotes, including humans.

Subsequent studies (Shu et al. 2007; Szollosi and Derenyi 2009) have challenged some of Borenstein and Ruppin’s results, although they have confirmed the finding that the mutational robustness of natural pre-miRNAs is higher than that of random sequences with the same structure. However, previous studies on the evolution of pre-miRNA robustness have two important limitations (Bonnet et al. 2004; Borenstein and Ruppin 2006; Shu et al. 2007; Shu et al. 2008; Szollosi and Derenyi 2009). First, they assume that random or shuffled sequences provide adequate null models for the evolution of pre-miRNAs, whereas natural sequences tend to evolve over much shorter sequence distances (Ehrenreich and Purugganan 2008; Liang and Li 2009; Nozawa et al. 2010). Second, the pre-miRNAs considered are not phylogenetically independent (Felsenstein 1985). Here we use a rigorous phylogenetic framework (Figure 2.1) to test whether or not the mutational and environmental robusteness of *Drosophila* pre-miRNAs have been subject to selection during 60 million years of evolution.

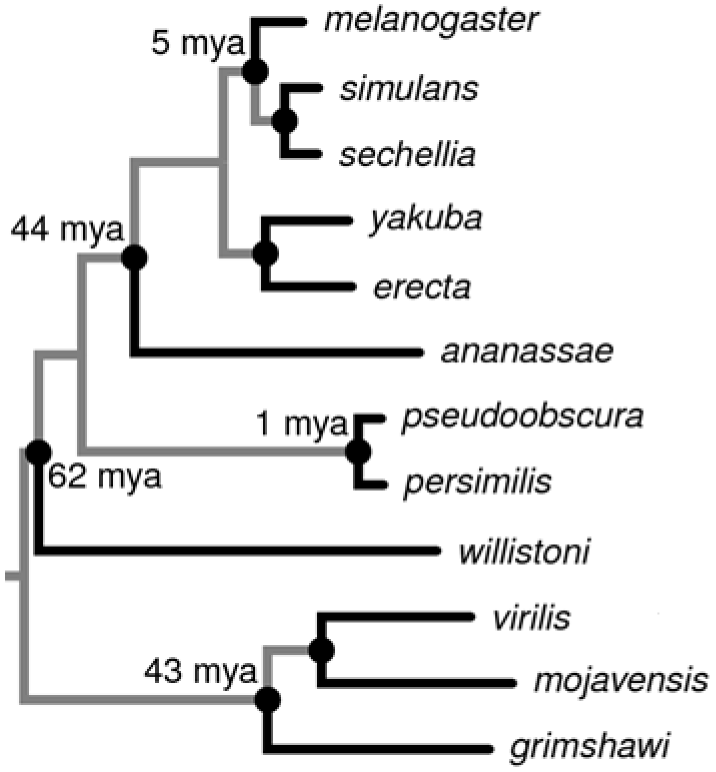

Figure 2.1: Orthologous pre-miRNA genes from *Drosophila* were analyzed using the above phylogenetic tree. (Siepel et al. 2005; Rosenbloom et al. 2010). (See http://tinyurl.com/drostree for original.) Divergence dates were taken from (Tamura et al. 2004). Only evolutionary events that occurred on the tips (black lines) were counted. For genes found in all twelve species, inferred ancestors were used at the nodes with black circles.

Materials and Methods

Assembly of orthologous pre-miRNA genes and ancestral sequence reconstruction

Orthologous *Drosophila* pre-miRNAs were downloaded from miRBase version 14 (Sept. 2009) (Griffiths-Jones et al. 2008). If a pre-miRNA gene had orthologs in at least 8 of the 12 *Drosophila* species found in miRBase (Figure 2.1), we included it for ancestral sequence reconstruction. If a species had multiple copies of a gene, we excluded all copies in that species.

We gathered a total of 71 pre-miRNA orthologous gene sets and aligned the sequences for each gene using MAFFT v6.717b (globalpair/G-INS-i alignment algorithm with default parameters and maximum iterations at 1000) (Katoh and Toh 2008). The guide tree used for the alignments was the phylogenetic tree in Figure 2.1 (Siepel et al. 2005; Rosenbloom et al. 2010). When a gene had orthologs in fewer than 12 species, the tree was pruned to remove the missing OTUs. To reconstruct ancestral sequence states, we used the web server ANCESTORS v1.0 (http://ancestors.bioinfo.uqam.ca/ancestorWeb/), which implements a maximum likelihood method (Blanchette et al. 2008; Diallo et al. 2010). Ancestry was inferred from our alignments and guide trees using the "best exact scenario" option and default parameters. Ancestral state reconstruction did not take into account the secondary structures of the sequences involved. We restricted our analyses to the terminal branches in the ancestral reconstruction that included at least one substitution and no insertions or deletions (indels), resulting in 221 usable branches.

Secondary structure prediction and simulation of alternative descendants

For each of the 221 terminal branches included in our analysis, we predicted the minimum free energy (MFE) structure of the ancestor and descendant using the folding algorithm developed by Zuker and Stiegler (1981) and implemented in the VIENNA RNA package version 1.8.4 (Hofacker et al. 1994). We then simulated alternative descendants for each branch by randomly mutating the ancestral sequence based on the number of substitutions in the natural descendant and keeping sequences that had the same structure as our descendant sequence (Zuker and Stiegler 1981). The number of possible descendants that are k substitutions away from an ancestor of sequence length L is $\binom{L}{k}3^k$ (assuming no back mutations). Because this number quickly becomes very large, we exhaustively searched all possible descendants for branches that contained 1 or 2 substitutions. For branches that contained $k \geq 3$ substitutions, we uniformly generated random descendants with replacement. For these searches, the sampling algorithm stopped when either it found 1,000 descendants with the same MFE structure (a success) or the probability of finding a descendant with the same MFE structure was less than 10^{-6}. To estimate this probability we used pseudocounts:

$$(2.1) \qquad P(\text{same structure}) = \frac{S+1}{N+2}$$

where S is the number of successes and N is the total number of sequences tried.

In addition to simulating possible descendants with the same MFE structure, we also simulated possible descendants without constraining on structure. As before, we uniformly generated random descendants of the ancestors of each sequence in our 221-branch dataset. However, we simply kept the first 1,000 simulated descendants for any value of *k*. Because structure was not constrained, these sets contained some sequences with the same MFE structure as the natural descendant and some with a different structure. We refer to the two sets of simulations as structure-constrained and structure-unconstrained, respectively.

Measuring robustness

Robustness is best measured as a variance (Wagner et al. 1997; Rice 1998; Gibson and Wagner 2000), but the robustness metrics used in previous studies of pre-miRNAs (Borenstein and Ruppin 2006; Shu et al. 2007; Szollosi and Derenyi 2009) do not capture this principle. (Note, however, that employing the metrics defined in those studies does not change my results qualitatively.) Here we introduce variance measures of robustness based on the base-pair distance (*d*) between two structures calculated in the VIENNA RNA package (Hofacker et al. 1994) (the number of base pairs present in one structure, but not the other).

We define the mutational “fragility” of a sequence of length *L* as

(2.2)
$$f_m = \frac{1}{3L}\sum_{i=1}^{3L}\left(\frac{d_i}{L}\right)^2$$

where d_i is the MFE structural distance between the sequence and its mutant neighbor *i*. This statistic is inversely related to robustness ($f_m = 0$ for a maximally robust sequence).

We measure mutational robustness as $r_m = 1 - f_m$. We define the environmental fragility of a sequence of length L as the variance of its structural ensemble:

$$f_e = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{2}\left(\frac{d_j}{L}\right)^2 \tag{2.3}$$

where d_j is the distance between a sampled pair of structures from the ensemble, and N is the number of sampled pairs. As before, environmental robustness (or thermodynamic stability) is calculated as $r_e = 1 - f_e$. We generated ensembles via VIENNA RNA's partition function folding algorithm using the default temperature of 310 K, and calculated f_e from $N = 10^6$ sampled pairs.

Drosophila pre-miRNA trees

Reconstruction of the 71 pre-miRNA genes produced a total of 813 terminal branches, approximately half had no changes, a quarter had indels, and a quarter had only substitutions. The number of terminal branches with only substitutions was 221. About half of these branches contained a single substitution, but 18% had 4 or more substitutions allowing us to explore a range of evolutionary divergence values (Table 2.1).

		Number of Substitutions						
		1	2	3	4	5	6+	Total
Size of Null Distribution	0–19	36	13	0	1	1	4	55
	20–99	75	11	0	0	0	0	86
	100–999	1	12	1	0	4	5	23
	1000+	0	11	21	12	8	5	57
	Total	112	47	22	13	13	14	221

Table 2.1: Table 1. Number of substitutions and size of structure-constrained null distributions per branch

Two methods of simulation were used to generate structure-constrained null distributions for each of these 221 branches, which produced a range of sample sizes for these null distributions (Table 2.1). Because some branches were exhaustively searched and others were randomly sampled, some null distributions may have duplicated sequences. Eighty branches had at least 100 samples, and 86 had between 20 and 99 samples. We further pruned these 221 branches by estimating the mutational and environmental robustness values of each of the samples in their null distributions and excluding branches that had less than 20 *unique* mutational robustness values. This produced a final dataset that contained 165 branches.

Test of selection

To determine whether pre-miRNA sequences have been selected for increased robustness, we compared the robustness of the natural sequences to null distributions produced in our simulations. Significance was assessed by first calculating the quantile *q* of each natural descendant's *r* value in the null distribution provided by the set of simulated descendants selected to have the same MFE structure. Because ties between *r* values were possible, we calculated *q* as the mid-point of any *r* values in the null-distribution that were the same as the descendants:

$$q = \frac{\sum_j I\left(r_j < r\right) + \tfrac{1}{2}\sum_j I\left(r_j = r\right)}{N} \tag{2.4}$$

where *r* is the robustness value (mutational or environmental) for the natural descendant, r_j is the value for the *j*-th element of the null sample, *N* is the number of values in the null sample for that branch, and we is an indicator function. If the natural descendants were not systematically selected for robustness, then we would expect them to follow their associated null distributions, so that the values of *q* should be uniformly distributed.

To evaluate the uniformity of the distribution of *q*, we used the Anderson-Darling goodness-of-fit test (Anderson and Darling 1952; Marsaglia and Marsaglia 2004). The Anderson-Darling test statistic, *A*, is based on the area between a sample cumulative distribution function (CDF) and the diagonal (the uniform CDF):

$$A = -n - \frac{1}{n}\sum_{k=1}^{n}\left(2k-1\right)\ln\left[x_k\left(1-x_{n+1-k}\right)\right] \tag{2.5}$$

where $x_1 < x_2 < \cdots < x_n$ is an ordered set of samples (Marsaglia and Marsaglia 2004). If this statistic is greater than expected, then the sample is considered to deviate significantly from uniformity. The significance of *A* was measured with the statistical software R (R Development Core Team 2009) using the method by Marsaglia and Marsaglia (2004) implemented in the ADGofTest package. Because our dataset contained uneven sample sizes and ties, we confirmed the significance levels via simulation. We constructed 1000 simulated datasets by randomly sampling robustness values from the null distribution of each of our branches and calculating the *A* statistic for each dataset.

Under the null hypothesis, the CDF consists of uniform order statistics, which follow a beta(*k*, *n*–*k*+1) distribution, where *k* is the rank of a point, and *n* is the sample size. From the null distribution, we determined the 95% concentration band for the simultaneous and equal-tail test of points in the CDF. We found that a pointwise concentration band with α = 0.000925 rejected only 5% of uniform Monte Carlo simulations with 165 points.

Results

Mutational and environmental robustness have not increased

If a substantial fraction of *Drosophila* pre-miRNAs have experienced a recent history of selection (direct or indirect) for increased robustness, then we might expect descendant (extant) pre-miRNAs to be more robust than their ancestors. When we compare the 221 descendants and their predicted ancestors, we find that both mutational robustness (r_m)

and environmental robustness (r_e) have decreased slightly (median $\Delta r_m = –0.0015\%$ and $\Delta r_e = –0.0031\%$; Figure 2.2) and that the change is marginally statistically significant (paired Wilcoxon test: $p_m \approx p_e \approx 0.05$). These results suggest that *Drosophila* pre-miRNAs have *not* evolved increased mutational and/or environmental robustness.

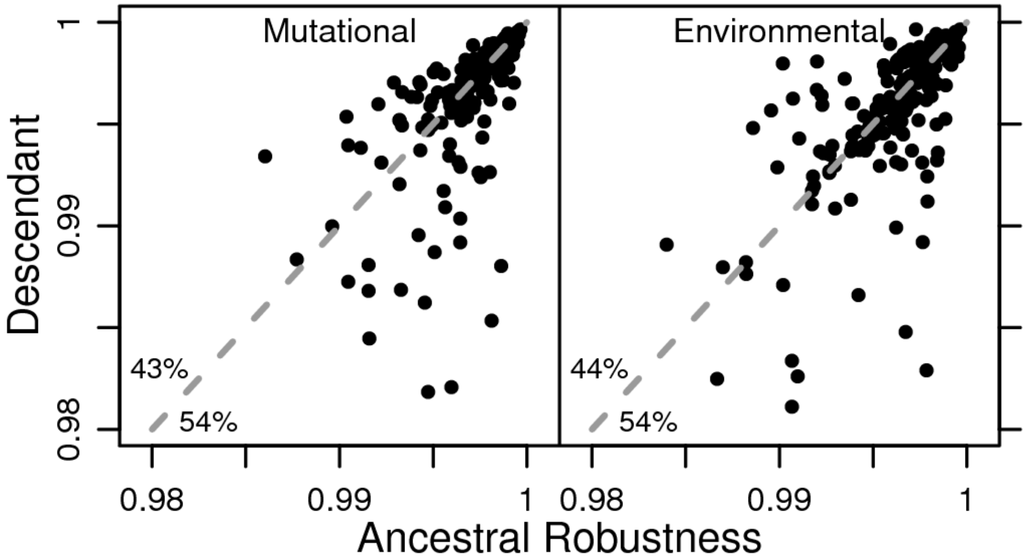

Figure 2.2: Mutational and environmental robustness have not increased from ancestor to descendant. Comparisons of mutational and environmental robustness in the 221 natural pre-miRNAs used in this study with estimated values in corresponding ancestors. Lines are $y = x$ diagonals. Numbers in the bottom left of panels indicate the percentage of points in either half of the panel.

Neutral evolution of both mutational and environmental robustness

The previous test assumed that neutral evolution of robustness would cause robustness not to change (on average) between ancestor and descendant. However, this assumption would not be met if, for example, most mutations caused a reduction in pre-miRNA robustness. To take such a possibility into account we replayed the tape of pre-miRNA evolution (Gould 1989); we generated a null distribution of descendant pre-miRNA sequences at the same *sequence* distance (k) from the ancestor and with the same secondary *structure* as the real descendant (i.e., with base-pair distance, $d = 0$). Figure 2.3 shows the resulting null distribution for mutational robustness for a representative sequence, the *D. pseudoobscura* mir-317. The null distribution allows us to measure the extent to which a real descendant pre-miRNA is more or less robust than expected under neutral evolution, when structure is the only constraint. For example, dps-mir-317 corresponds to the $q = 54.7\%$ quantile of the null distribution, implying that it is slightly more robust than expected, despite being slightly *less* robust than its ancestor (Figure 2.3). If the robustness of *Drosophila* pre-miRNAs has been evolving neutrally, then we expect that values of q over the entire dataset should be uniformly distributed.

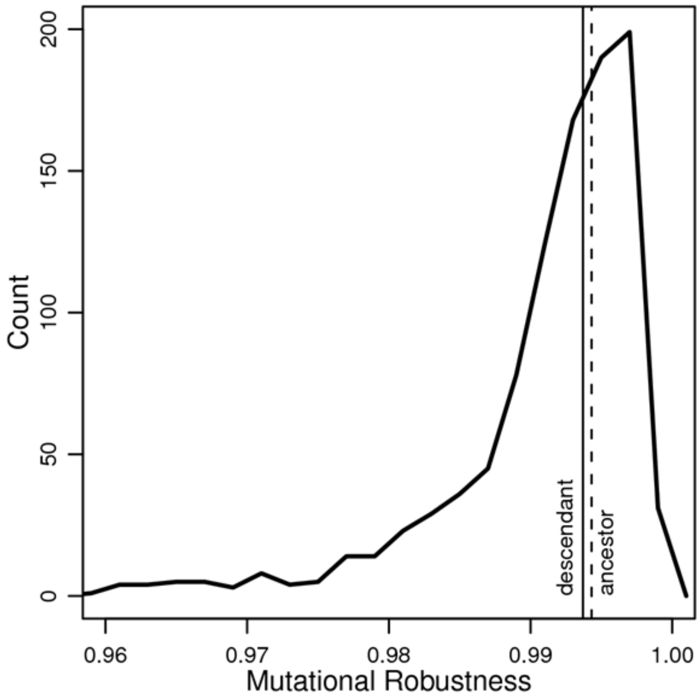

Figure 2.3: Density plot of the structure-constrained null distribution of mutational robustness for dps-mir-317. The null distribution consists of 1000 sequences differing from the ancestor in any k = 5 nucleotide positions but with exactly the same length (L = 90 nt) and structure as dps-mir-317.

The cumulative distribution functions (CDFs) of q for mutational and environmental robustness are plotted in Figure 2.4A and do not differ significantly from a uniform distribution (Anderson-Darling test: $p_m = 0.469$ and $p_e = 0.480$, $N = 165$). The insignificance of these goodness-of-fit tests was confirmed by simulation ($p_m = 0.463$ and $p_e = 0.481$; Fig. 2.4B). These results suggest that the robustness of *Drosophila* pre-miRNAs has evolved neutrally.

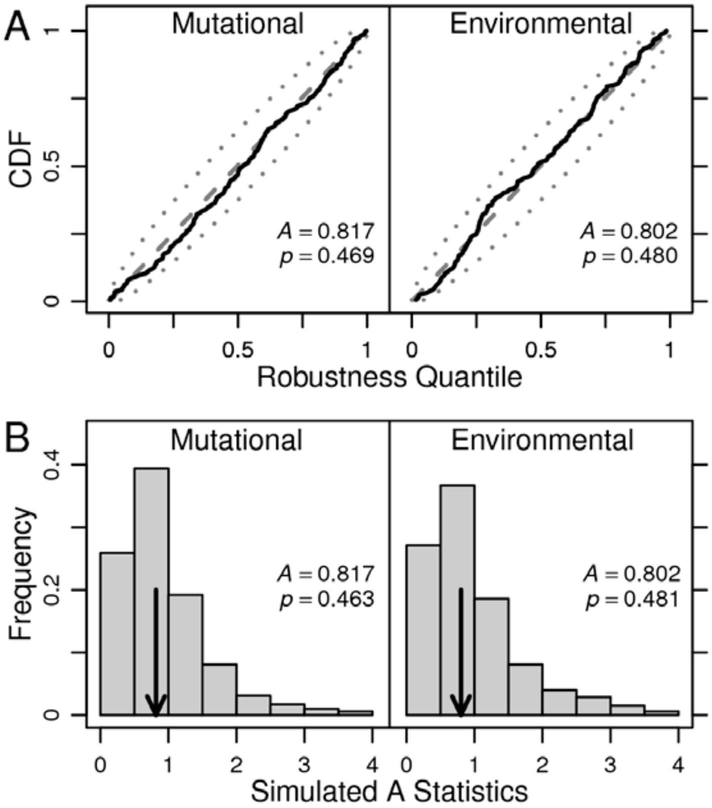

Figure 2.4: Neutral evolution of robustness. (A) The cumulative distribution function (CDF) of *q* values (black lines) of the robustness of pre-miRNAs compared to their corresponding structure-constrained null distributions. Anderson-Darling test statistics (*A*) and their associated *p*-values are also shown. The dashed lines represent the expected values of points in a CDF for a uniform distribution, and dotted lines mark 95% concentration bands (only 5% of uniform CDFs of this size are expected to have at least one point outside this region). (B) Histograms of the simulated distributions for each *A* statistic; *p*-values do not change noticeably.

Strong purifying selection against changes in secondary structure

So far, we have imposed an absolute constraint on the structure of the natural pre-miRNAs, as did earlier studies (Borenstein and Ruppin 2006; Shu et al. 2007; Szollosi and Derenyi 2009). Does this assumption make a difference for the outcome of our analysis? To test this assumption, we generated a new null distribution of descendant pre-miRNA sequences. These were at the same *sequence* distance (k) from the ancestor as the real descendant, but their *structure* was not constrained in any way (i.e. we allowed any value of d between real and simulated descendant). We then repeated the analysis described in the previous section. The CDFs of q are plotted in Figure 2.5A and show a highly statistically significant deviation from a uniform distribution (Anderson-Darling test: $p < 4 \times 10^{-6}$ for both r_m and r_e): ~75% of descendants are more robust than expected ($q > 0.5$).

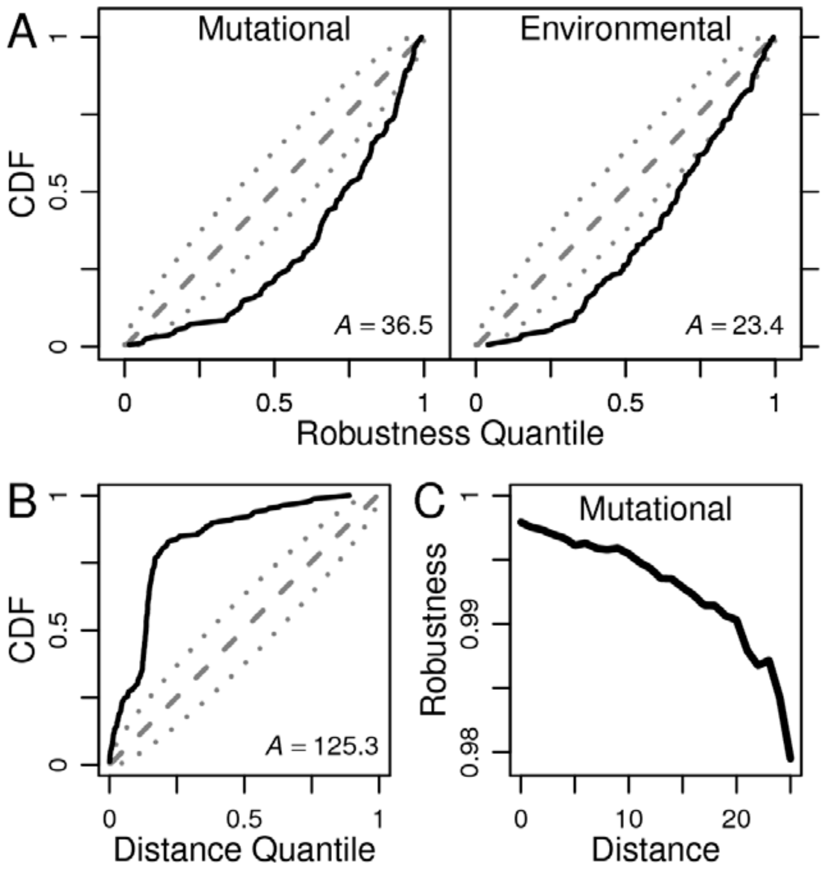

Figure 2.5: Strong purifying selection against changes in secondary structure. (A) CDF of *q* values of the robustness of pre-miRNAs compared to their corresponding structure-unconstrained null distributions. (B) CDF of *q* values of the structural distance between ancestor and descendant pre-miRNAs compared to their corresponding structure-unconstrained null distributions. (C) Median mutational robustness of the sequences from the structure-unconstrained null distributions binned according to their structural distance from the natural descendant pre-miRNA. Robustness decreases as the distance to natural structures increases.

These results are caused by variation in structure; that is, both mutational and environmental robustness tend to decrease as the structures of simulated sequences deviate more from the structure of the corresponding (natural) descendant pre-miRNA (Figure 2.5C). This result indicates that the constraint on structure is a crucial assumption of these analyses. What might cause such a constraint? One possibility is that there is strong purifying selection against all mutations altering the pre-miRNA structure. If so, then the structures of descendants should be closer to those of their ancestors than expected by chance. To test this prediction, we used the structure-unconstrained null distribution of descendant pre-miRNA sequences and employed the same approach we used for robustness in the previous section. Over 90% of descendants were, indeed, structurally closer to their ancestors than expected under neutral evolution (q > 0.5; Anderson-Darling test: $p < 4 \times 10^{-6}$; Figure 2.5B).

Therefore, the evolution of *Drosophila* pre-miRNAs is consistent with the operation of strong purifying selection in which the functional constraint is the secondary structure and both mutational and environmental robustness are evolving neutrally.

Discussion

Theoretically, the strength of direct selection for mutational robustness is at most the magnitude of the deleterious mutation rate (Kimura 1967; Proulx and Phillips 2005); thus, direct selection for mutational robustness should not operate on the pre-miRNAs of multicellular eukaryotes. Against this expectation, Borenstein and Ruppin (2006) concluded that eukaryotic pre-miRNAs are under direct selection for mutational robustness. We investigated the 60-million-year evolutionary history (Tamura et al. 2004) of mutational and environmental robustness of *Drosophila* pre-miRNAs. We replayed the tape of pre-miRNA evolution based on several explicit evolutionary models. Our analyses provided no evidence that either kind of robustness evolved under any form of selection.

Our conclusion, like those from earlier studies (Borenstein and Ruppin 2006; Shu et al. 2007; Szollosi and Derenyi 2009), postulates the existence of a strong constraint on pre-miRNA structure. We have shown that one plausible mechanism for this constraint—strong purifying selection—can explain the observed pattern of evolution in secondary structure (Figure 2.5). Strong purifying selection can also account for the observation that *Drosophila* pre-miRNAs evolve ~30% slower than nonsynonymous sites of protein-coding genes (Nozawa et al. 2010). Indeed, there is strong evidence that pre-miRNAs are subject to stringent structural constraint: the precise structure of a pre-miRNA influences several aspects of its maturation including recognition and cleavage by Drosha and nuclear export by Exportin 5 (Zeng and Cullen 2003, 2004, 2005; Zeng et al. 2005; Han et al. 2006).

Selection against changes in pre-miRNA structure may indirectly result in mutational and environmental robustness. Borenstein and Ruppin (2006) claimed that the pre-miRNAs of multicellular eukaryotes show signs of direct selection for mutational robustness. Our results refute this claim and show that inverse folding produces invalid null distributions, e.g., for mir-317-dps, inverse folding of random sequences (Borenstein and Ruppin 2006; Szollosi and Derenyi 2009) is equivalent to simulating alternative descendants without constraining for structure (Figure 2.6).

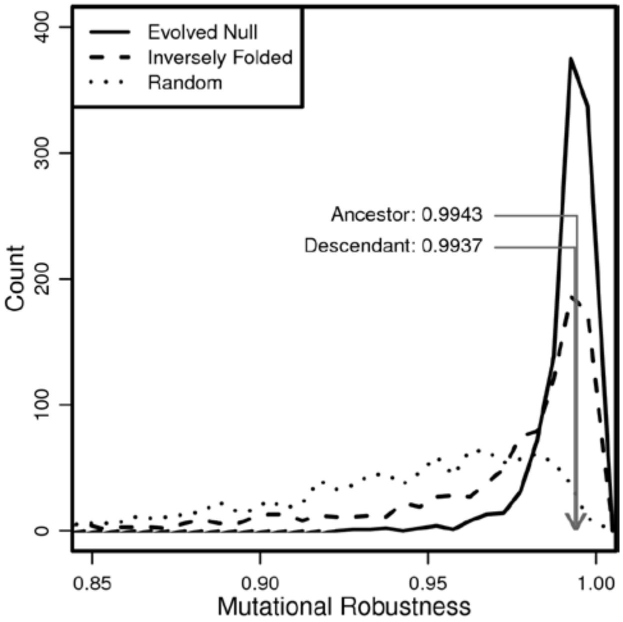

Figure 2.6: Density plot of different null distributions of mutational robustness for dps-mir-317. Evolved null: sequences differing from the ancestor in any $k = 5$ nucleotide positions but with exactly the same structure as dps-mir-317 (structure constrained). Inversely folded: random sequences evolved through a random walk until their secondary structure matches that of dps-mir-317 exactly (Borenstein and Ruppin 2006). Random: random sequences with any structure.

We also find no evidence that *Drosophila* pre-miRNAs have experienced direct selection for mutational robustness in the last 60 million years. This is in agreement with theoretical expectations. If we assume $L = 95$ and $P_{del} = 85\%$ (Nozawa et al. 2010), then we predict that $2N_eU \approx 0.60 < 1$ in *D. melanogaster* (Lynch and Conery 2003). Under these conditions, direct selection for mutational robustness would be ineffectual.

Despite the clear advantages of our approach over those employed in earlier studies, it does have three limitations. First, we only consider a single high-likelihood reconstruction of the evolutionary history of each orthologous gene. The uncertainty involved in ancestral state reconstruction could be incorporated into these analyses through Bayesian phylogenetic methods like that of Robinson et al. (2003). Second, we allowed nucleotide substitutions to occur anywhere in a sequence, when it is clear that different regions of *Drosophila* pre-miRNA sequences evolve at different rates (Nozawa et al. 2010). Third, we did not consider indels, when they have obviously played an important role in pre-miRNA evolution. However, there is no reason to assume that these limitations have biased our analyses.

In conclusion, contrary to earlier claims, there is no evidence that pre-miRNAs have been shaped by direct selection for any kind of robustness.

Chapter 3: Identification of orthologous processed pseudogenes in mammals

Abstract

Processed pseudogenes are formed through the reverse transcription of mature mRNAs. Because they are considered “dead on arrival” and to evolve neutrally, they can be used as a null model to test for selection. Part of my research involved using orthologous processed pseudognes to test for selection on synonymous sites. In this Chapter, I describe the procedures used to identify processed pseudogenes and corresponding orthologs, as well as discuss my results.

Introduction

A pseudogene is a nongenic DNA segment that exhibits a high degree of similarity to a functional gene, but contains defects such as nonsense and frameshift mutations that prevent it from being properly expressed (Graur 2014). There are three general types of pseudogenes; duplicated, unitary, and processed. However, due to segmental duplications (Bailey and Eichler 2006) we can also have duplication of the above pseudogenes. Duplicated pseudogenes are formed when a functional gene is duplicated, and subsequently one of the copies acquires mutations that render it nonfunctional. These sequences usually retain the characteristics of genes, including a promoter region and an exon-intron structure. Unitary pseudogenes are formed when a single copy gene acquires a mutation that prevents it from being properly translated, or transcribed, and therefore becomes nonfunctional (Graur 2014). Finally, processed pseudogenes are formed through the reverse transcription of the mature mRNA of a gene (Esnault et al. 2000). After reverse transcription, the cDNA is randomly inserted back into the genome through the action of an endonuclease (Esnault et al. 2000). Because processed pseudogenes are derived from a mature mRNA product they lack the upstream promoter of a normal gene; therefore they are considered “dead on arrival”, becoming non-functional immediately after formation (Graur 2014).

There are four main characteristics of processed pseudogenes: (1) they lack introns and promoters; (2) have a poly-A-tail at the 3’ end; (3) their flanking regions consist of direct repeats that are associated with insertion sites of transposable elements (Rouchka and

Cha 2009) ; and (4) because they are derived from a mature mRNA they show sequence similarity to the cDNA of their “parent” gene.

Using some of the above features, studies have identified processed pseudogenes in a variety of species, including human and mouse (Zhang et al. 2002; Ohshima et al. 2003; Torrents et al. 2003; Zhang and Gerstein 2004; van Baren and Brent 2006). In human and mouse, the number of processed pseudogenes identified ranged from 7,000 to 9,000 (Ohshima et al. 2003; Zhang and Gerstein 2004; van Baren and Brent 2006).

In the present study, using a modified version of Zhang et al’s (2002) method, I identified processed pseudogenes in a set of thirteen high coverage mammalian genomes. These pseudogenes were then used to identify orthologs between different pairs of species. The current Chapter provides a detailed description of the procedures used, and compares the methods and results of the current study to Zhang et al.’s.

Materials and Methods

Identification of processed pseudogenes

The genomes and cDNA sequences of 13 mammalian species (human, chimp, gorilla, macaque, marmoset, mouse, rat, rabbit, pig, cow, dog, horse) were downloaded from the ENSEMBL 64 database (Flicek et al. 2012). To identify processed pseudogenes I applied a modified version of the method by Zhang et al. (2002). The procedure is depicted in Figure 3.1 and details of each step are explained below:

(1). *BLASTN search and initial filtering of blast hits.*

For each species, the set of cDNAs was blasted against each chromosome using the BLASTN program (Altschul et al. 1990). Hits were retained only if they had a significant e-value ($<10^{-9}$). The hits that remained consisted of both genes and pseudogenes. Using the coordinates of annotated exons in ENSEMBL I removed hits that overlapped with exons and were left with hits to pseudogenes. A large fraction of remaining hits were from query cDNAs contaminated by repeats, low complexity regions, and transposable elements. Repeats and low complexity regions can cause false positive hits, while the insertion of transposable elements in exonic regions, usually causes non-functionalization (Abrusan et al. 2008). To identify and remove such cases from our dataset I used the program RepeatMasker (Smit 1996-2012).

(2). *Further analysis of hits corresponding to a query cDNA*

After removing dubious hits, I further analyzed each query and its corresponding hits. During the first step I identified cases in which a genomic region showed significant similarity to two or more regions within the query cDNA. Such cases may represent duplications of exonic regions or low complexity regions. Because only one of the regions could be the homologous region I chose the one with the lowest e-value and as an alternative criterion the one with the highest similarity. In the second step I identified cases in which a query cDNA had multiple neighboring hits. I only regarded instances where hits were within 1000 nucleotides of each other. The space between hits may represent introns of duplicated pseudogenes, insertions within pseudogenes, or, regions between tandemly duplicated pseudogenes. If neighboring hits showed similarity to the same cDNA region they were identified as tandem duplications and therefore regarded as separate entries in the database. However, if the hits corresponded to different regions within the cDNA, and were in the right orientation, both query and genomic coordinates were merged. Such cases were regarded as duplicated pseudogenes or processed pseudogenes disrupted by insertions.

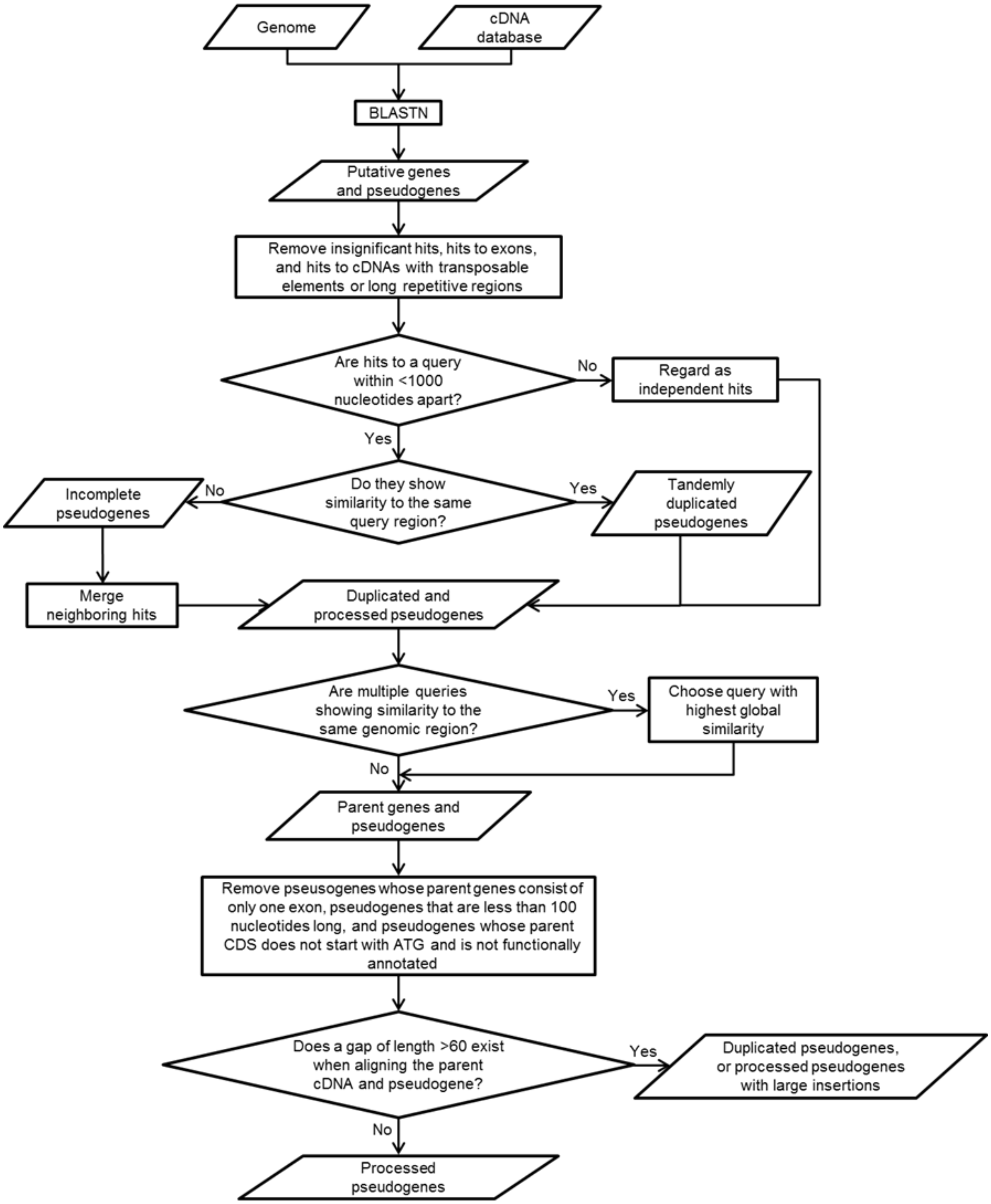
Genome
cDNA database
BLASTN
Putative genes and pseudogenes
Remove insignificant hits, hits to exons, and hits to cDNAs with transposable elements or long repetitive regions
Are hits to a query within <1000 nucleotides apart?
No
Regard as independent hits
Yes
Do they show similarity to the same query region?
No
Incomplete pseudogenes
Yes
Tandemly duplicated pseudogenes
Merge neighboring hits
Duplicated and processed pseudogenes
Are multiple queries showing similarity to the same genomic region?
Yes
Choose query with highest global similarity
No
Parent genes and pseudogenes
Remove pseusogenes whose parent genes consist of only one exon, pseudogenes that are less than 100 nucleotides long, and pseudogenes whose parent CDS does not start with ATG and is not functionally annotated
Does a gap of length >60 exist when aligning the parent cDNA and pseudogene?
Yes
Duplicated pseudogenes, or processed pseudogenes with large insertions
No
Processed pseudogenes

Figure 3.1: A flow chart depicting the procedure used to identify processed pseudogenes. The procedure included: (1) BLASTing a species set of cDNAs against its genome and removing dubious hits. (2) Merging neighboring hits belonging to a query. (3) Removing

short pseudogenes and identifying parent genes. (4) Removing duplicated pseudogenes or processed pseudogenes with large insertions.

(3). *Removing short pseudogenes and identifying parent genes*

At this point in our analysis the set of pseudogenes consisted of a large number of short sequences. These short sequences may represent heavily fragmented pseudogenes, short reverse transcriptions, or regions showing similarity just by chance. Such sequences were removed from the dataset by retaining only those that covered at least 1/3 of the query coding sequence. Some of the remaining pseudogenes showed significant similarity to cDNAs of more than one gene. In such cases, in order to identify which gene gave rise to the pseudogene (known as the "parent gene"), I first aligned each cDNA to the pseudogene using the global alignment method MAFFT (Katoh et al. 2002) and then estimated the percent similarity. The gene with the cDNA showing the highest similarity was regarded as the parent gene. To further ensure that the parent genes in our dataset were functional and not wrongly annotated pseudogenes I removed entries in which the coding sequence of the parent gene did not start with ATG, and was not functionally annotated by ENSEMBL. In most organisms the most common start codon is ATG with alternate start codons (non ATG) being rare (Kozak 1999). A coding sequence that does not start with an ATG initiation codon may represent one of the rare cases, or maybe a gene that recently became a pseudogene.

(4). *Filtering out duplicated pseudogenes*

The remaining steps of our procedure were aimed at removing duplicated pseudogenes and retaining processed pseudogenes. To do so I used two important criteria: i) pseudogenes had to show similarity to more than one exon within the query cDNA and ii) when aligning the cDNA to the pseudogene, gaps present in the cDNA, had to be less than sixty in length. The former criterion was applied to remove pseudogenes created from single exon genes, or ones created from a single exon within a gene. Such pseudogenes can only be verified as processed if they have a clear poly-A-tail (Zhang et al. 2003) or target site duplications (Terai et al. 2010). These features however, are quickly lost in sequences that are nonfunctional and evolving neutrally. The latter criterion was applied to remove duplicated pseudogenes. Introns present in duplicated pseudogenes form gaps when aligned to the query cDNA. Considering that almost all introns in mammals are much larger than 60 bp (Zhang et al. 2003; Pozzoli et al. 2007) pseudogenes with smaller insertions are most likely processed pseudogenes.

Identification of orthologous processed pseudogenes

For each of the processed pseudogenes in mouse and human I extracted the sequence that corresponded to the coding sequence (CDS) of the parent gene. To identify orthologous pseudogenes in primates and rodents I only used processed pseudogenes that included at least 100 nucleotides of the CDS region. These included 6931sequences in human and 9750 in mouse. The human processed pseudogenes were used to identify orthologous processed pseudogenes in chimpanzee, orangutan, macaque and marmoset and the mouse pseudogenes were used to identify orthologs in rat. To achieve the above goal I followed the procedure in Figure 3.2. Details of each step are described below:

(1-3). The human processed pseudogenes were blasted against each of the primate genomes and the mouse processed pseudogenes against the rat genome using BLASTN (Altschul et al. 1990). Genome sequences were downloaded from the ENSEMBL database (Flicek et al. 2012). The resulting hits were filtered according to their e-value and percent similarity. Specifically, hits with an e-value $>10^{-9}$ and below a specified similarity (<90% in great apes, <80% in macaque, <75% in marmoset and <60% in rat) were removed. To further remove random hits I filtered out short sequences. Specifically, I retained hits whose lengths were at least 70% of the corresponding query when comparing human to great apes, 60% in human-macaque and human-marmoset comparisons, and 50% in mouse-rat. The differences in percentages were based on the different levels of divergence between species. Finally, I removed hits that overlapped annotated exonic regions.

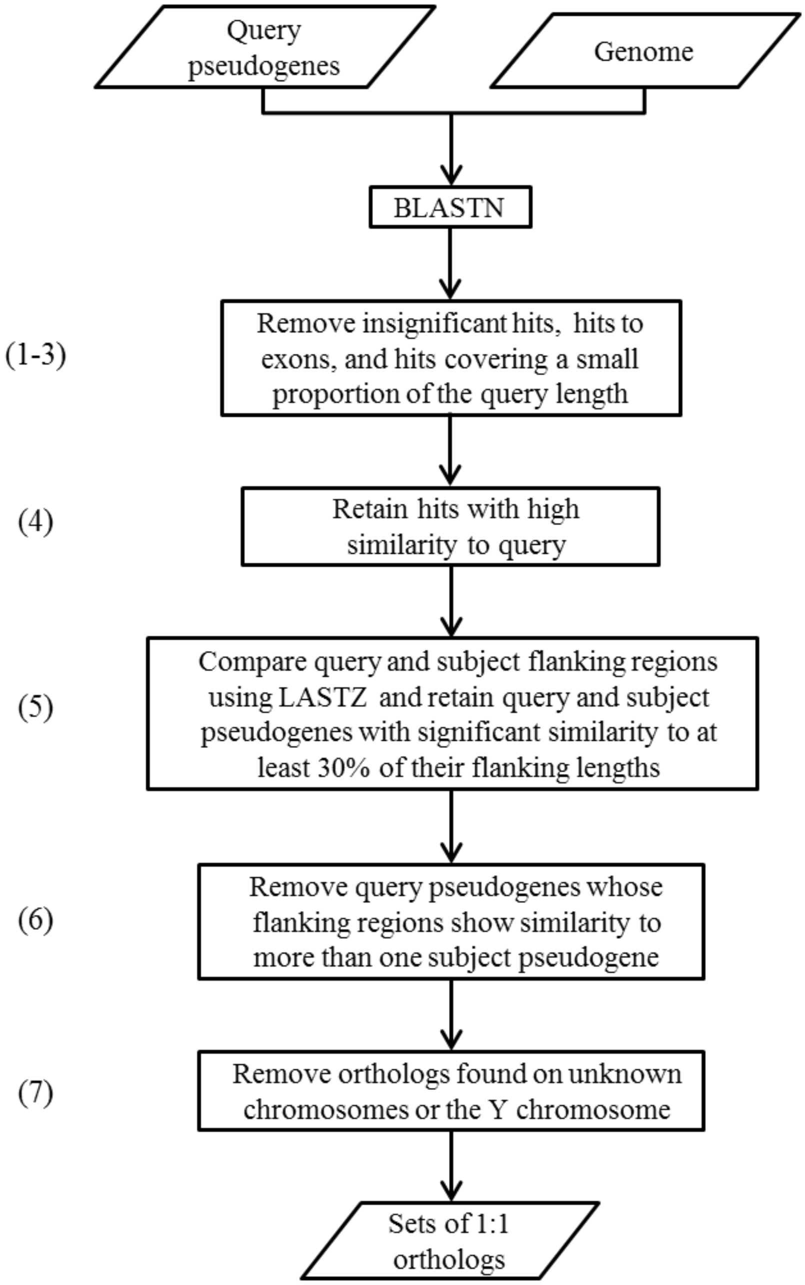
Query pseudogenes
Genome
BLASTN
(1-3)
Remove insignificant hits, hits to exons, and hits covering a small proportion of the query length
(4)
Retain hits with high similarity to query
(5)
Compare query and subject flanking regions using LASTZ and retain query and subject pseudogenes with significant similarity to at least 30% of their flanking lengths
(6)
Remove query pseudogenes whose flanking regions show similarity to more than one subject pseudogene
(7)
Remove orthologs found on unknown chromosomes or the Y chromosome
Sets of 1:1 orthologs

Figure 3.2: A flowchart depicting the procedure taken to identify orthologous processed pseudogenes. The numbers in parenthesis correspond to the numbers in the text.

(4). The remaining pseudogenes were then ranked according to similarity to the query pseudogene. Orthologous pseudogenes should show the highest similarity between each other. To gather candidate orthologous pseudogenes I chose ones that shared a very high similarity. Specifically, if x% was the highest similarity, hits that had a similarity of ≤ x-3% in the primate datasets and ≤ x-5% in the rodent data set were kept for further analysis.

(5). After all of the above steps were performed, the remaining number of hits in each data set was: 9749 human-chimpanzee, 9984 human-orangutan, 34367 human-macaque, 26834 human-marmoset and 52535 in mouse-rat. The next step, involved comparing flanking regions of the query pseudogenes (human and mouse) to the flanking regions of the corresponding subject pseudogenes (chimp, orangutan, macaque, marmoset and rat). This step should further narrow down the list of possible orthologs because processed pseudogenes are randomly inserted in the genome and therefore their flanking regions should most of the times share significant similarity with their orthologous counterparts. To perform the above step, I fetched 2000 nucleotides upstream and downstream when comparing human to great apes and old world monkeys, and 5000 for the human-marmoset and mouse-rat data sets. I used longer flanking regions in the latter cases because of high divergence. The flanking regions in each of the six data sets were compared using LASTZ (Harris 2007). LASTZ is a local alignment tool that was developed to align long genomic regions between highly divergent species. Like BLAST it first finds local regions of significant similarity and then increases the sensitivity to see whether these local regions can be extended into longer syntenic alignments. The resulting hits were then filtered according to the length of the flanking regions that showed significant similarity. Only hits that covered at least 30% of the flanking length were kept for further analysis.

(6). After applying the above step the number of possible orthologs that remained was: 6192 human-chimpanzee, 5243 human-orangutan, 3107 human-macaque, 1332 human-marmoset and 1827 mouse-rat. Due to duplicated processed pseudogenes these sets included query pseudogenes that showed similarity to more than one pseudogene. To avoid paralogous processed pseudogenes I removed such cases resulting in 4971 human-chimpanzee, 4535 human-orangutan, 3107 human-macaque, 1332 human-marmoset and 1827 mouse-rat 1:1 orthologous processed pseudogenes.

(7). Finally, I removed cases in which processed pseudogenes were not assigned to a chromosome, or in the case of human, the pseudogene was found on the Y-chromosome. The latter case represents instances where human pseudogenes on the Y-chromosome were orthologous to pseudogenes on the X-chromosome. These pseudogenes were located in pseudoautosomal regions which are known to undergo crossing over (Charchar et al. 2003). The above filtering resulted in a final data set of 4961 human-chimpanzee, 4507 human-orangutan, 3107 human-macaque, 1332 human-marmoset and 1827 mouse-rat orthologous processed pseudogenes.

Results

The number of processed pseudogenes identified in each species is depicted in Figure 3.3. Processed pseudogenes were identified using cDNAs as query sequences. Column A corresponds to the total number of processed pseudogenes identified for each species. Column B depicts the numbers of pseudogenes that consist of at least 100 nucleotides of the query coding region. The mean length of the processed pseudogenes is ~700 nucleotides and the median length is 539.

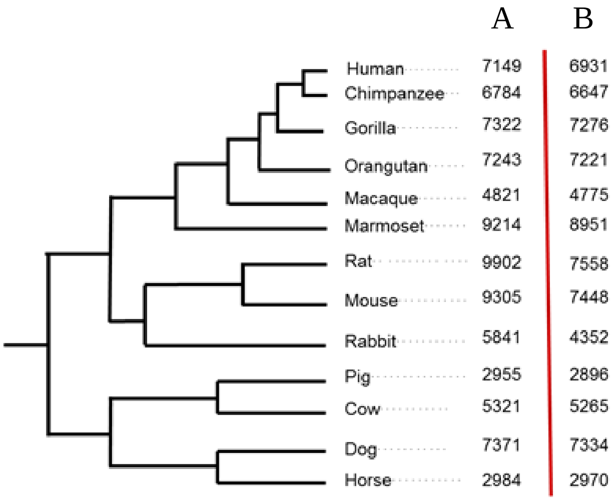

Figure 3.3: The number of processed pseudogenes identified for each species. Column A corresponds to the total number of processed pseudogenes identified and column B corresponds to the number of pseudogenes that show similarity to at least 100 nucleotides of the query CDS region.

74% of the processed pseudogenes consist of frameshift mutations when aligned to their parent CDS. Furthermore, 64% of them do not start with a start codon due to mutations or 5’ truncations. When comparing the set of human processed pseudogenes to the set downloaded from pseudogene.org (Karro et al. 2007) there were 4,273 processed pseudogenes identified by both studies, 2,876 that were unique to the present study, and 4,507 unique to pseudogene.org (Figure 3.4).

The 4,507 pseudogenes missing from the final dataset, were not included because of the following reasons: 565 were located on patches (patches are additional sequences for alternate alleles that were not present on the primary assembly or additional sequences that will replace misassembled regions in the next assembly); 635 had parent genes with signals of transposable elements; 1701 covered only one exon of the parent gene; and 1606 were identified as duplicated.

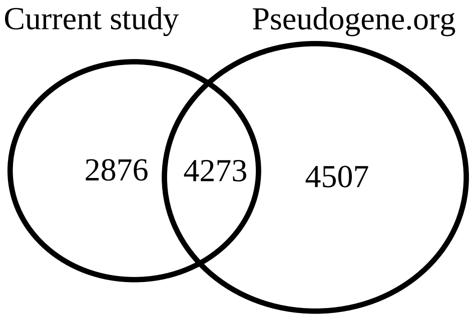

Figure 3.4: A Venn diagram depicting the number of common and unique human processed pseudogenes annotated by pseudogene.org and identified by the present study.

Figure 3.5 shows the number of human genes with a corresponding number of processed pseudogenes. The majority of genes give rise to a single processed pseudogene, with a very few number producing more than two processed pseudogenes (Figure 3.5).

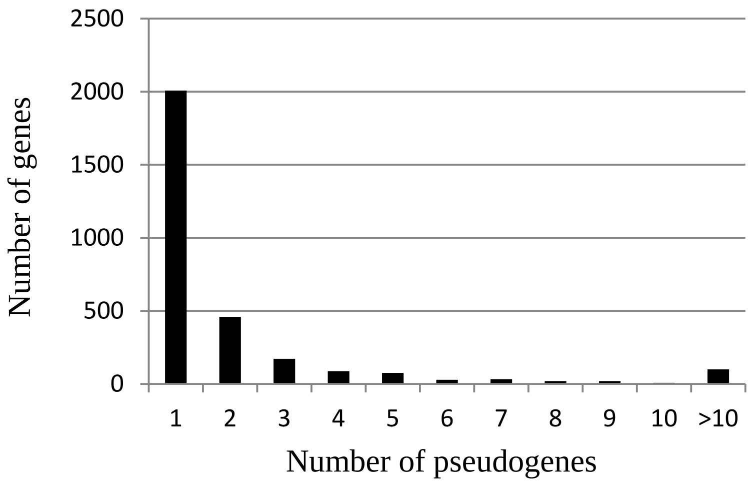

Figure 3.5: A histogram showing the number of human genes with a specified number of processed pseudogenes. For example, ~2000 genes gave rise to one processed pseudogene.

Table 3.1 shows the mean pairwise divergence between othologous pseudogenes estimated using the F84 model in PHYLIP (Felsenstein 1989) and Figure 3.6 depicts the distributions of these pairwise distances. The distances were measured for orthologous processed pseudogenes in which there were at least 100 aligned nucleotides after masking poorly aligned regions.

Table 3.1: Mean pairwise divergences between orthologous pseudogenes ($\bar{d}$).

	Number of orthologs	$\bar{d}$
human-chimp	3985	0.013
human-orangutan	3544	0.033
human-macaque	2474	0.064
human-marmoset	1068	0.108
mouse-rat	1391	0.158

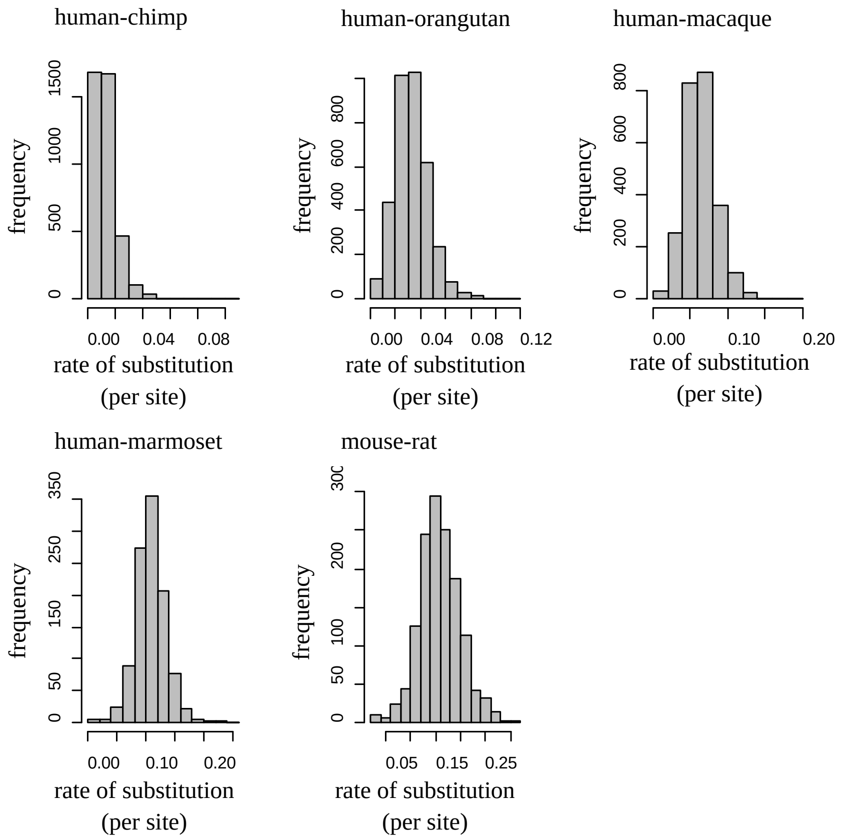

Figure 3.6: Histograms of the pairwise divergences between orthologous processed pseudogenes.

Discussion

Using cDNA sequences as query sequences I identified processed pseudogenes in 13 high coverage mammalian genomes (Figure 3.3). The method applied was a modified version Zhang et al.’s method (2002) which is implemented by the database of pseudogenes, pseudogene.org (Karro et al. 2007). Despite sharing some common steps, the method implemented in the present study differed in some important ways. Zhang et al. used protein sequences as queries and TBLASTN to identify local similarities with processed pseudogenes. TBLASTN aligns protein sequences to a nucleotide database translated in all six frames. Given that processed pseudogenes do not have an intact reading frame and do not evolve under any selective constraints this may result in misalignments.

```
ACT   GCA      CDS
AC—   ACA      pseudogene
```

For example, given that the CDS above is read using frame 1, and in the pseudogene there is a deletion at position three and a change from a G to an A at position four, TBLASTN may mistakenly align ACT with ACA since both code for threonine. Another important difference is that when using protein sequences as queries one cannot identify introns in untranslated regions. Finally, in the study by Zhang et al. they used similarities estimated from local alignments to identify the parent genes for each pseudogene, while in the present study I use global alignments. Because local similarity does not reflect the overall similarity of a gene to its pseudogene this may result in wrongly identifying the parent gene.

After comparing the human processed pseudogenes identified by the present study to the ones downloaded from pseudogene.org (Karro et al. 2007) ~60% were common to both data sets (Figure 3.4) however there were a large number that were unique to each study. The majority of pseudogenes that were unique to pseudogene.org were not included in the final data set because they originated from single exon genes, the parent cDNA consisted of regions common to transposable elements and finally they were identified as duplicated. I did not include processed pseudogenes to single exon genes because in these cases it is hard to distinguish duplicated from processed pseudogenes. In addition, I did not include pseudogenes whose parent cDNAs were predicted to show similarity to transposable elements, to ensure that the parent gene is functional. Finally, processed pseudogenes that were identified as duplicated by the current study formed large gaps at untranslated regions when aligned to the parent cDNA. These gaps could be caused by introns present in duplicated pseudogenes.

When examining the number of processed pseudogenes originating from each parent gene (Figure 3.5) it seems that the majority of genes give rise to a single pseudogene with a very smaller number having more than two pseudogenes. Studies have shown that highly expressed genes produce a higher number of processed pseudogenes (Harrison et al. 2005). However, a higher number may also result from tandem duplications of processed pseudogenes (Khurana et al. 2010).

Assuming that processed pseudogenes are under no selective constraint they should accumulate frameshift mutations. After aligning processed pseudogenes to their parent CDSs, 74% of the processed pseudogenes contained frameshift mutations. Furthermore, 64% of the processed pseudogenes did not contain the 5’ translation initiation codon, because of 5’ truncations or mutations.

In addition to identifying processed pseudogenes in mammals I identified pairs of orthologs in primates and rodents (Table 3.1). Species belonging to the laurasiatheria superorder were excluded because early analysis showed that sequences in this group of species were too divergent. Figure 3.6, shows the distributions of pairwise distances between orthologs. The small number of outliers observed in great apes and old world monkeys may result from random errors in pairwise distances or wrongly identified orthologs. After removing these outliers, the resulting sets of orthologous processed pseudogenes were used in Chapter 4 to estimate selection at synonymous sites.

Chapter 4: No evidence for purifying selection on synonymous mutations in mammals

Abstract

Selection on synonymous sites in mammals has been highly controversial with no definite conclusion. In the present study using processed pseudogenes as the neutral model of evolution I test for selection on synonymous sites in primates and rodents. After establishing that processed pseudogenes are evolving neutrally, I test for selection on synonymous sites by comparing the pairwise synonymous divergence between genes and pseudogenes. Unlike previous estimates in which synonymous sites were found to evolve at ~70% the rate of corresponding sites in pseudogenes, our results indicate that synonymous sites are evolving at very close to the neutral expectation (≥ 92%). With the use of simulated sequences I show that the small deviation from neutrality observed in the present study, and the large difference previously obtained, could result from the shift in GC content after pseudogene formation.

Introduction

The question of whether synonymous sites are under selection is of great importance because the d_N/d_S ratio used to detect selection in protein coding genes, assumes that synonymous sites evolve neutrally. This assumption however is being challenged, because numerous studies indicate possible selection on synonymous sites (Smith and Hurst 1998; Ophir et al. 1999; Bustamante et al. 2002; Hellmann et al. 2003; Pagani and Baralle 2004; The Chimpanzee Sequencing and Analysis Consortium 2005; Doherty and McInerney 2013). The proportion of synonymous sites estimated to be under selection, ranges from 1-39% (Ophir et al. 1999; Bustamante et al. 2002; Hellmann et al. 2003; Parmley et al. 2006; Eory et al. 2010; Caceres and Hurst 2013).

The functions found to be associated with synonymous sites, and which could explain the signature of selection, are: (1) sysnonymous sites may harbor signals for the splicing machinery; (2) synonymous sites play a role in mRNA stability, (3) serve as binding sites for miRNAs and transcription factors (Hurst 2006; Gu et al. 2012; Stergachis et al. 2013), and (4) are part of codons that are more efficient and accurate during translation (Ikemura 1985; Akashi and Eyre-Walker 1998; Wright et al. 2004; Stoletzki and Eyre-Walker 2007; Drummond and Wilke 2008)

The most frequently used methods to estimate selection on synonymous sites fall into two categories: (1) assessing codon usage bias (Akashi and Eyre-Walker 1998; Duret 2002; Wright et al. 2004; Zhou et al. 2010; Kiyohara et al. 2012; Lawrie et al. 2013) and (2) measuring the difference in evolutionary divergence between synonymous sites, and regions that are assumed to be nonfunctional and to evolve neutrally (e.g., introns, ancestral repeats and processed pseudogenes) (Smith and Hurst 1998; Ophir et al. 1999; Bustamante et al. 2002; Hellmann et al. 2003; Comeron 2006; Eory et al. 2010).

Evidence of codon bias caused by selection against codons that are translationally insufficient has been primarily observed in species with large effective population sizes such as *Drosophila melanogaster, Saccharomyces cerevisiae, Arabidopsis thaliana,* and *Caenorhabditis elegans* (Duret 2002; Wright et al. 2004; Zhou et al. 2010; Lawrie et al. 2013). In mammals, evidence of selection for translational efficiency has been weak (Kanaya et al. 2001; Lander et al. 2001; Chamary et al. 2006). However, it has been shown that selection against protein misfolding can explain most variation in codon usage (Drummond and Wilke 2008).

Most studies comparing the divergence of nonfunctional regions (e.g., introns, ancestral repeats, and processed pseudogenes) to the divergence between synonymous sites, estimate different levels of selection. Specifically, results range from 1-39 % of synonymous sites being under selection (Ophir et al. 1999; Bustamante et al. 2002; Hellmann et al. 2003; Eory et al. 2010). Using divergences between orthologous genes, it was found that about 1-9 % of synonymous sites at exon-intron boundaries are more conserved than others (Parmley et al. 2006; Caceres and Hurst 2013). These regions were suggested to function as exon splice enhancers (ESE), and aid in the accurate splicing out of introns to form the mature mRNA.

In the present study, I estimate selection on synonymous sites in a variety of taxa, including, great apes, monkeys, and rodents. To estimate selection on synonymous sites, I use the set of processed pseudogenes identified in Chapter 3. After confirming that the majority of pseudogenes evolved neutrally I compare rates of evolution at synonymous sites in orthologous genes (d_{S_f}) to the rates of corresponding sites in pseudogenes (d_{S_ψ}). Assuming that synonymous sites are under no selection, the ratio d_{S_f}/d_{S_ψ} should be approximately equal to 1. In addition to using orthologous genes and pseudogenes I estimated the ratio d_{S_f}/d_{S_ψ} using the approach presented by Bustamante et al. (2002). Finally, using simulated data I assessed how the estimates of $d_{S_f}$and d_{S_ψ} are affected by changes in GC content.

Materials and Methods

Identification processed pseudogenes and their corresponding orthologs

The method used to identify processed pseudogenes and corresponding orthologs is explained in Chapter 3.

Identifying the parent genes of the orthologous processed pseudogenes

The parent genes of human processed pseudogenes were used to identify orthologs in chimp, orangutan, macaque, and marmoset, while, mouse parent genes were used to identify orthologs in rat. Orthologs were downloaded from ENSEMBL 72 (Flicek et al. 2013). Parent CDSs were downloaded for the majority of orthologous processed pseudogenes (4,118 human-chimp, 3,723 human-orangutan, 2,606 human-macaque, 1,075 human-marmoset, 1,606 mouse-rat). In a few cases there were more than one possible candidate parents (107 human-chimp, 15 human-orangutan, 351 mouse-rat). To identify the parent CDS in such cases, the processed pseudogenes were aligned with their possible parent CDSs using MAFFT (Katoh et al. 2002). Poorly aligned regions were filtered out using the evaluation mode in T-COFFEE (Notredame et al. 2000), and then the percent similarity was estimated between the processed pseudogene and each possible parent. The CDS with the highest similarity was chosen to be the parent.

Alignment of processed pseudogenes and parent genes by codon positions

Orthologous processed pseudogenes and their corresponding orthologous parent CDSs were aligned and low quality alignment regions were masked. After alignment, the reading frame of a functional coding sequence was used. Gaps in the functional sequences were regarded as insertions in the pseudogenes, and gaps in the pseudogenes as deletions in the processed pseudogenes. Codons were disregarded if there was a corresponding deletion in the pseudogene, they contained poorly aligned nucleotides, or if they represented a stop codon. For further analyses, I only kept alignments that contained at least 100 nucleotides. The resulting data set included the following numbers of codon alignments: 3,941 (human-chimp), 3,483 (human-orangutan), 2,414 (human-macaque), 1,055 (human-marmoset), 1,372(mouse-rat). These sets are not mutually exclusive but consist of human orthologs that are shared by more than one set.

Model of gene and pseudogene evolution

The model used to analyze rates of synonymous and nonsynonymous substitution in genes $\left(d_{N_s}, d_{S_s}\right)$ and pseudogenes (d_{N_ψ}, d_{S_ψ}) was developed by Yang and Nielsen (1998) and implemented in the PAML phylogenetic analysis package (Yang 1997). This treats a DNA sequence as a series of codons and specifies the substitution rate from codon *i* to *j* as

$$q_{ij} \propto \begin{cases} 0, & \text{if codons } i \text{ and } j \text{ differ at more than one codon position,} \\ \pi_j, & \text{for synonymous transversion,} \\ \kappa\pi_j, & \text{for synonymous transition,} \\ \omega\pi_j, & \text{for nonsynonymous transversion,} \\ \omega\kappa\pi_j, & \text{for nonsynonymous transition} \end{cases}$$

where κ is the transition-transversion ratio, ω is the d_N/d_S ratio, and π_j is the equilibrium frequency of codon *j*. π_j is estimated from the nucleotide frequencies at the three codon positions (see Yang and Nielsen 1998 for details) while the set of free parameters estimated by maximum likelihood is $\{t, \omega, \kappa\}$ where t is the expected number of nucleotide substitutions per codon. From the maximum likelihood estimates of t, ω, and κ one can estimate the number of substitutions per synonymous site (d_S) and the number of replacement substitutions per replacement site (d_N) (Yang and Nielsen 1998).

Testing the neutral evolution of processed pseudogenes

To identify whether a pair of orthologous processed pseudogenes are evolving neutrally, I compared the log likelihood of the model in which ω_ψ could be less or equal to one to the log likelihood of the model in which ω_ψ was equal to one. Assuming that the former model represents the alternative hypothesis (H_1) and the latter model the null hypothesis (H_0), I can test if ω_ψ is significantly different from 1 by considering that two times the difference between the log-likelihoods of the two models Ln_1-Ln_0 to be asymptotically distributed as a χ^2 random variable with one degree of freedom. To further test whether there is a significant difference in the rate of substitution between nonsynonymous ($\overline{d_{N_\psi}}$) and synonymous ($\overline{d_{S_\psi}}$) sites in pseudogenes I used a bootstrap approach. Specifically, for 10,000 bootstrap replicates I estimated the difference between the means $\overline{d_{N_\psi}}$ and $\overline{d_{S_\psi}}$. Using the bootstrap distributions I estimated the 95% confidence interval for the ratio $\overline{d_{N_\psi}}/\overline{d_{S_\psi}}$. A 95% confidence interval that included one indicated no significant difference. The final test used to estimate whether pseudogenes evolve neutrally is to estimate the ratios d_{ψ_1}/d_{ψ_2} and d_{ψ_1}/d_{ψ_3}. d_{ψ_1}, d_{ψ_2}, d_{ψ_3} are the pairwise divergences between the first, second, and third "codon" positions in orthologous processed pseudogenes. Assuming, that processed pseudogenes are evolving neutrally these ratios should be approximately equal to one. Divergences were estimated using the F84 model in PHYLIP (Felsenstein 1989).

Simulating the evolution of orthologous processed pseudogenes

I used the program DAWG (Cartwright 2005) to simulate sequences similar to the processed pseudogenes used in our study. For each pair of orthologous pseudogenes (Table 1) I used half the value of the pairwise divergence as branch lengths in our simulations. Pairwise divergence was estimated using the F84 model in PHYLIP (Felsenstein 1989). As root sequences I used the human parent CDS in the primate simulations and the mouse parent CDS in the rodent simulations. This captured the GC content at synonymous sites. The transition-transversion ratio was set at $\kappa = 4$ as it is observed in mammals (Rosenberg et al. 2003). To simulate the change in GC content of processed pseudogenes after formation I set equilibrium nucleotide frequencies at A = 0.30, G = 0.20, C = 0.20, T = 0.30. Indels were not included in the simulation. For all other parameters used in the simulation I used the default settings. After removing stop codons from the simulated sequences, $\overline{d_N}$ and $\overline{d_S}$ were estimated using CODEML (Yang 1997).

Annotated and simulated sequences for comparing methodologies

This section outlines the procedure used to assemble a set of annotated and simulated sequences that were later used to compare our results with those obtained by Bustamante et al. (Bustamante et al. 2002).

To assemble sets of sequences, similar to those used by Bustamante et al (Bustamante et al. 2002) I followed the procedure explained below. First I gathered a subset of 2162 human pseudogenes in which their parent genes had a corresponding ortholog in the mouse or marmoset genome. Orthologous relations were downloaded using the BioMart tool available at Ensembl 72 (Flicek et al. 2013). Alignments of the human pseudogenes, the parent genes and the mouse or marmoset orthologs were built, and regions of poor alignment quality were removed. After the formation of codon alignments as explained in the above section I were left with 1990 alignments. For each codon alignment, using the *free-ratios* model implement in the program *codeml,* I estimated the parameters $\{t_f, t_o, t_\psi, \omega_f, \omega_o, \omega_\psi, \kappa\}$. o represents the branch leading to the outgroup species, ψ is the branch leading to the pseudogene and f denotes the functional paralog (Figure 4.1).

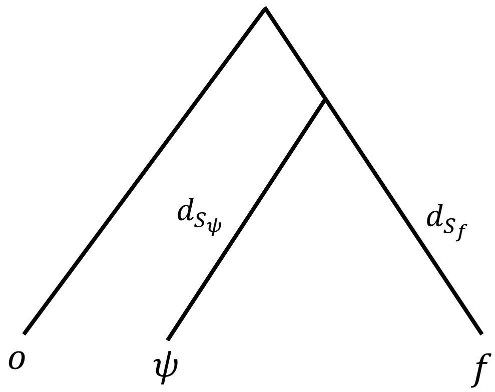

Figure 4.1: In the above phylogeny f depicts the codon sequence of a functional gene, ψ represents the codon sequence of a pseudogene, and o represents the codon sequence of the outgroup. The rate of synonymous substitution was estimated over the branches leading to the parent gene d_{S_f} and the corresponding processed pseudogene d_{S_ψ}.

The above seven parameter model was compared to a nested six parameter model that assumed $\omega_\psi = 1$. Sequences that rejected the null hypothesis $\omega_\psi = 1$ at the 5% significance level were removed. Furthermore I removed data sets where $t_o < t_\psi$, or if any of the branch lengths was greater than one. The former condition assured that the pseudogenes were formed after speciation, and the latter removed highly diverged sequences. Our final datasets included 959 codon alignments in which mouse was the outgroup species, and 157 in which marmoset was the outgroup. Estimates of d_{S_f} and d_{S_ψ}were only kept for sequences with at least 100 synonymous sites in order to minimize random fluctuations. These included 572 estimates where mouse was the outgroup, and 104 in which marmoset was the outgroup.

The simulations carried out to test the methodology applied by Bustamante et al. (2002) followed a scenario in which mouse or marmoset were the possible outgroups and the corresponding parent gene and pseudogene were human (Figure 4.2).

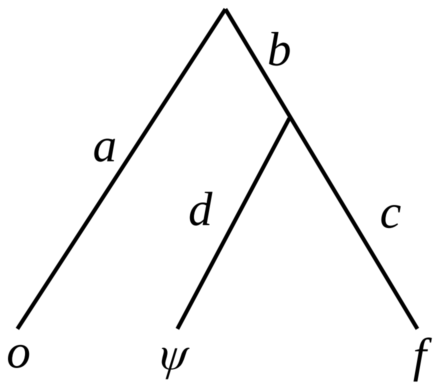

Figure 4.2: The phylogeny used for simulating the formation of processed pseudogenes after the human-mouse and human-marmoset split. *o, ψ and f* represent the outgroup, pseudogene, and parent gene, respectively. The abbreviations *a*, *b*, *c* and *d* represent the branches manipulated during the simulations.

For each scenario I performed three simulations in which the internal branches *d* and *c* were equal and had the following values: 0.01, 0.03 and 0.05. In the case where mouse was the outgroup $a = 0.35$ and $b = \{0.14, 0.12, 0.10\}$. These branches were chosen so that $b + c = 0.15$. The values correspond to the approximate divergence between mouse and human which is ~0.5. In the case of marmoset $a = 0.07$ and $b = \{0.04, 0.02, 0.01\}$. These values correspond to an approximate divergence between human and marmoset of about 0.12-0.13. These values were based on the 46 species tree built by the ucsc genome browser using fourfold degenerate sites (http://genomewiki.ucsc.edu/index.php/Human/hg19/GRCh37_46-way_multiple_alignment)

Using the codon model by Nielsen and Yang (Yang and Nielsen 1998) which is implemented in INDELIBLE (Fletcher and Yang 2009), I simulated the sequences of the outgroup genes (o), parent genes (f), and the ancestral sequences of the genes and pseudogenes (Figure 3.1). Gene simulations were done using a transition/transversion ratio (κ) of 2.5 and 4.0. The value of $\kappa = 4$ is observed in mammals (Rosenberg et al. 2003) while $\kappa = 2.5$ is the default value for INDELIBLE. The selection values (ω) used were 0.1 and 0.2. are similar to those observed in mammals (Eyre-Walker et al. 2002; The Chimpanzee Sequencing and Analysis Consortium 2005). In order to simulate sequences with a similar GC content as in the real sequences I used the codon frequencies estimated by a pairwise comparison of the 959 human-mouse gene orthologs in our real data set. Codon frequencies were estimated by CODEML (Yang 1997). Because the branch lengths mentioned above are expressed as number of synonymous substitutions per synonymous site I had to transform them to number of substitutions per codon for the simulation of genes. These can be derived by the equation $t = (3\ d_S\ \times\ \rho_S^1)/\rho_S^*$ where t is the number of substitutions per codon, d_S is the number of substitutions per synonymous sites, ρ_S^1 is the proportion of synonymous mutations before selection, and ρ_S^* is the proportion of synonymous substitutions after selection (Yang and Nielsen 1998, 2000). The above equation can be rewritten as $t = 3\ d_S \times (\rho_S^1 + \omega \times \rho_N^1)$ where ω is the d_N/d_S ratio and ρ_N^1 is the proportion of nonsynonymous mutations before selection (Yang and Nielsen 1998, 2000). To simulate the corresponding processed pseudogenes (ψ) (Figure 3.1) I used the program DAWG (Cartwright 2005) and the ancestral sequence of the gene and pseudogene generated by INDELIBLE. For the GTR model implemented in DAWG I used the same κ values as in the gene simulations and the equilibrium base frequencies for G, C, A, and T were set at 0.2, 0.2, 0.3, 0.3 respectively. These frequencies reflect the approximate GC content of the human genome (~40%). Processed pseudogenes were also simulated using the codon model implemented in INDELIBLE. In the first set of simulations gene and pseudogene equilibrium codon frequencies were set to be equal and in the second set of simulations pseudogene equilibrium codon frequencies were estimated using the base frequencies G=0.2, C=0.2, A=0.3, and T=0.3. For these simulations $\omega = 1$. In all simulations performed indels were excluded, and the length of the sequences produced was 1200 nucleotides or 400 codons.

To simulate the orthologous genes and pseudogenes used to test the method in the present study I simulated sequences whose pairwise distances $d_{S_{\psi_{ab}}}$ and $d_{S_{f_{ab}}}$(Figure 4.3) were 0.01, 0.03, 0.06, 0.1 and 0.17. These values captured the typical pairwise divergences between the orthologous pseudogenes used in the present study.

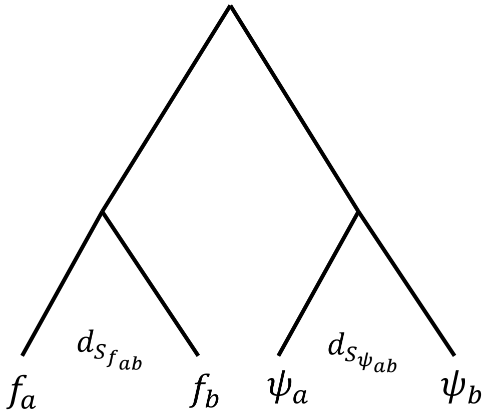

Figure 4.3: In the present study using othologous pseudogenes (ψ_a, ψ_b) and their genes (f_a, f_b) I estimated the rates of synonymous substitution $d_{S_{\psi_{ab}}}$ and $d_{S_{f_{ab}}}$.

For each pairwise distance I performed 959 simulations using the same parameters as mentioned in the previous simulations. The root sequences produced by INDELIBLE during the simulation of genes were used by DAWG to simulate the corresponding pseudogenes. Equilibrium GC content was set at 40% and $\kappa = 4$

Bootstrap analysis

To test whether the rate of substitution at nonsynonymous sites in pseudogenes, $\overline{d_{N_\psi}}$, was significantly different than the rate at synonymous sites $\overline{d_{S_\psi}}$ I computed the 95% confidence of the ratio $\overline{d_{N_\psi}}/\overline{d_{S_\psi}}$. Using a paired bootstrap approach I sampled (d_{N_ψ}, d_{S_ψ}) with replacement and calculated $\overline{d_{N_\psi}}/\overline{d_{S_\psi}}$. Using 10,000 bootstrap samples I estimated the 95% confidence intervals for the above ratio. Each bootstrap replicate is a random sample of the same size as the original data set. A confidence interval that included one indicates no significant difference. The same procedure was used for all ratios examined in the study, including $\overline{d_{S_f}}/\overline{d_{S_\psi}}$ and $\overline{d_{S_{f_{ab}}}}/\overline{d_{S_{\psi_{ab}}}}$

Results

Do processed pseudogenes evolve neutrally?

Assuming that processed pseudogenes are under no selection, the ratio of "nonsynonymous" to "synonymous" substitutions should be approximately equal to one $\left(\omega_{\psi} = 1\right)$. To test the above assumption, for each pair of orthologs, I compared a model that allowed for selection to a model that assumed neutral evolution $\left(\omega_{\psi} = 1\right)$. Considering all species pairs examined the percentage of pseudogene pairs that rejected the null hypothesis $\left(\omega_{\psi} = 1\right)$ at the 5% significance level ranged from 6-9 %. As a second test I looked at the rate in the mean rate at nonsynonymous and synonymous sites $(\overline{d_{N_\psi}}/\overline{d_{S_\psi}})$ (Table 4.1). As indicated by the results in Table 4.1, on average, nonsynonymous sites evolve at a rate that is ~90% the corresponding rate at synonymous sites.

Table 4.1. Differences in the mean rate of substitution at nonsynonymous and synonymous sites in processed pseudogenes.

Species pairs	Number of Pairs	$\overline{d_{N_\psi}}/\overline{d_{S_\psi}}$ (95% CI)	$\overline{d_{N_S}}/\overline{d_{S_S}}$ (95% CI)
human-chimp	3571	0.92 (0.89, 0.95)	0.86 (0.83, 0.89)
human-orangutan	3225	0.89 (0.87, 0.91)	0.86 (0.84, 0.89)
human-macaque	2218	0.91 (0.89, 0.93)	0.90 (0.88, 0.92)
human-marmoset	1047	0.89 (0.87, 0.92)	0.87 (0.85, 0.89)
mouse-rat	1306	0.94 (0.92, 0.97)	0.92 (0.91, 0.94)

To examine whether the difference observed was due to a bias introduced by the method, I simulated a set of neutrally evolving sequences and then estimated the ratio ($\overline{d_{N_s}}/\overline{d_{S_s}}$) (Table 4.1). The ratio of the rate of nonsynonymous and synonymous substitutions is approximately the same between the simulated and pseudogene data (Table 4.1).

To whether the bootstrap distributions of the pseudogenes $\overline{d_{N_\psi}}/\overline{d_{S_\psi}}$ were significantly smaller than the corresponding distributions of the simulated sequences $\overline{d_{N_s}}/\overline{d_{S_s}}$ I estimated the probability a randomly chosen difference $\overline{d_{N_\psi}}/\overline{d_{S_\psi}}$ was larger than $\overline{d_{N_s}}/\overline{d_{S_s}}$. In all species comparisons except human-chimp and human-orangutan the results were insignificant. In the cases of human-chimp and human-orangutan I found the opposite, that is, the difference $\overline{d_{N_s}}/\overline{d_{S_s}}$ was significantly smaller than $\overline{d_{N_\psi}}/\overline{d_{S_\psi}}$. This difference may be result of the simulation model not fully capturing pseudogene evolution.

Do synonymous sites evolve neutrally?

In the present study, using sets of orthologous genes and pseudogenes (Figure 4.3) I estimated the average ratio $\overline{d_{S_{f_{ab}}}}/\overline{d_{S_{\psi_{ab}}}}$. Our results indicate that synonymous sites in genes and pseudogenes evolve at approximately the same rate (Table 4.2). In most species comparisons in which the mean rate of substitution at synonymous sites ($\overline{d_{S_{f_{ab}}}}$) was lower than the corresponding rate in pseudogenes ($\overline{d_{S_{\psi_{ab}}}}$) , the difference was only ≤8% (Table 4.2). $\overline{d_{S_{\psi_{ab}}}}$ and $\overline{d_{S_{f_{ab}}}}$ were also estimated over a set of simulated sequences to examine whether the slight but in some cases significant difference could be the result of a shift in GC content when processed pseudogenes are formed.

Table 4.2. The average ratio of substitution rates at synonymous sites in genes and pseudogenes. (N_s: number of synonymous sites between orthologs).

	Observed		Simulations
Species Pairs	Number of pairs $N_s \geq 100$	$\overline{d_{S_{f_{ab}}}}/\overline{d_{S_{\psi_{ab}}}}$(95% CI)	$\overline{d_{S_{f_{ab}}}}/\overline{d_{S_{\psi_{ab}}}}$(95% CI)
human-chimp	1458	1.00 (0.94, 1.06)	0.91 (0.86, 0.96)
human-orangutan	1446	0.91 (0.88, 0.95)	0.90 (0.87, 0.92)
human-macaque	885	0.92 (0.89, 0.96)	0.92 (0.90, 0.94)
human-marmoset	519	0.92 (0.89, 0.96)	0.92 (0.91, 0.94)
mouse-rat	409	0.97 (0.93, 1.00)	0.88 (0.87, 0.90)

Although, $d_{S_{\psi_{ab}}}$ and $d_{S_{f_{ab}}}$ were set to be equal to the neutral divergence between the species compared, $\overline{d_{S_{f_{ab}}}}/\overline{d_{S_{\psi_{ab}}}}$ was estimated to be less than one (Table 4.2). In some cases the ratios obtained from the simulated data were significantly different from the observed data (Table 4.2). These differences maybe caused because the parameters used in the simulations do not reflect the exact evolution of the genes and pseudogenes.

The effects of outgroup choice and equilibrium codon frequencies when estimating selection on synonymous sites

Using a similar approach to the one employed by Bustamante et al. (2002) I estimated the rate of substitution at synonymous sites in human genes (d_{S_f}) and pseudogenes (d_{S_ψ}) using mouse or marmoset as an outgroup species (Figure 4.1). The results in the mean ratio $\overline{d_{S_f}}/\overline{d_{S_\psi}}$ (Table 4.3) are separated into estimates obtained by a set of pseudogenes

Table 4.3. Comparing the rate of evolution at synonymous sites in genes and pseudogenes using different outgroups. (N_s: number of synonymous sites).

Outgroup	Number of sets $N_s \geq 100$	Ratio (95% CI)
mouse	572	0.80 (0.74, 0.86)
marmoset	104	0.90 (0.75, 1.08)
	Common set	
mouse	93	0.83 (0.69, 0.99)
marmoset		0.88 (0.72, 1.08)

that were formed after the human-mouse split and a smaller set formed after the human-marmoset split. $\overline{d_{S_f}}/\overline{d_{S_\psi}}$ estimates were also compared over a common set of 93 pseudogenes. The common set is less than 104 because in some cases when using mouse as the outgroup the number of synonymous sites were less than 100. In both cases, the ratio $\overline{d_{S_f}}/\overline{d_{S_\psi}}$ is significantly smaller than when using using mouse as an outgroup. Specifically, when using mouse as the outgroup $\overline{d_{S_f}}/\overline{d_{S_\psi}}$ the average estimate of the two sets is ~0.82, and when using marmoset it is ~0.89 (Table 4.3). The same pattern is observed when using simulated sequences (Figure 4.4). The results in Figure 4.4 are based on simulations in which pseudogene nucleotide or codon equilibrium frequencies were estimated using the human GC content of 40%. As shown in Figure 4.4, all types of simulations resulted in a $\overline{d_{S_f}}/\overline{d_{S_\psi}}$ ratio that was significantly

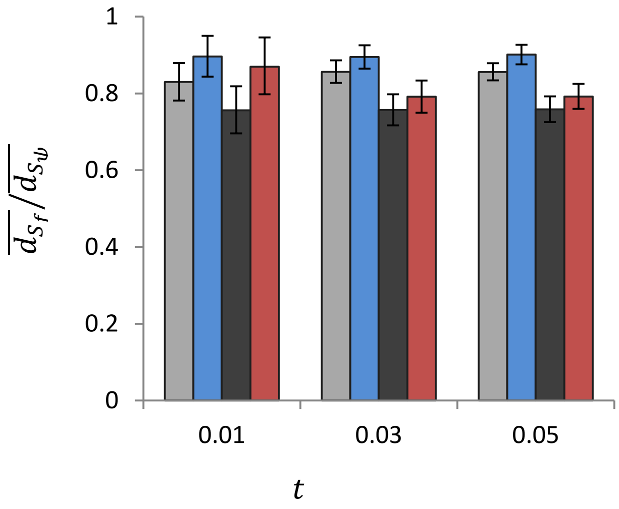

Figure 4.4: The mean $\overline{d_{S_f}}/\overline{d_{S_\psi}}$ ratio and 95% confidence intervals estimated from simulated sequences. t represents the branch length after reverse transcription and the units are in number of substitutions per site. Gray and blue columns represent the results from simulations obtained when marmoset was the outgroup. Gray columns represent nucleotide simulations and blue columns represent codon simulations. Black and red columns are the results obtained from simulations in which mouse was the outgroup. Black columns represent nucleotide simulations and red columns represent codon simulations. The results depicted in both figures were obtained using an $\omega = 0.2$ and $\kappa = 4.0$.

smaller than one. Deviations from the expected ratio of one increased when using a distantly related outgroup such as mouse (black and red columns). In addition to the previously mentioned simulations, I also performed a set of codon simulations in which equilibrium codon frequencies were the same along gene and pseudogene branches. In these cases $\overline{d_{S_f}}/\overline{d_{S_\psi}}$ was estimated to be equal to one (Data not shown).

Discussion

To examine whether synonymous mutations are under purifying selection I used processed pseudogenes as a model of neutral evolution. Although a few studies have shown that some processed pseudogenes are functional, and evolutionary conserved (Harrison et al. 2005; Frith et al. 2006; Svensson et al. 2006; Pei et al. 2012) the results of the present study indicate that in the majority of cases there is no evidence selection. The percentage of pseudogene pairs that rejected the null hypothesis $(\omega_{\psi} = 1)$ at the 5% significance level ranged from 6-9 %. These sequences may represent either false positives (type I error), or a set of retroposed genes that are under purifying selection (Emerson et al. 2004; Vinckenbosch et al. 2006). In addition to this result, the ratio $\overline{d_{N_{\psi}}}/\overline{d_{S_{\psi}}}$ is approximately equal to one (Table 4.1). Under the assumption that processed pseudogenes are non-functional, "nonsynonymous" and "synonymous" sites should evolve at the same rate. Although the ratio $\overline{d_{N_{\psi}}}/\overline{d_{S_{\psi}}}$ is slightly smaller than one the outcome of the simulations performed indicate that the difference in rate between $\overline{d_{N_{\psi}}}$ and $\overline{d_{S_{\psi}}}$ may stem from the change in GC content at "synonymous sites" after pseudogene formation (Table 4.1). Because processed pseudogenes are derived from the mature mRNA of a gene they are initially high in GC content. However, because they often land in regions of low GC content, they may experience a mutation pressure towards a GC content similar to their genomic neighborhoods.

After showing that the majority of processed pseudogenes do not show any evidence of selection, I estimated selection at synonymous sites using orthologous genes and pseudogenes (Table 4.2). The primate othologs depicted in Table 4.2 are not all unique but in the majority of cases share the same human orthologs.

The results of the present study indicate that synonymous sites in genes evolve at ~ 94% of the rate of “synonymous” sites in pseudogenes. This is almost equal to the neutral expectation. The slight difference observed could be the result of selection, or a shift in GC content experienced by processed pseudogenes. To examine the later cause, $\overline{d_{S_{f_{ab}}}}/\overline{d_{S_{\psi_{ab}}}}$ was estimated using simulated sequences in which $d_{S_{\psi_{ab}}}$ and $d_{S_{f_{ab}}}$ were set to be equal, but the equilibrium GC content of processed pseudogenes was set at the genomic average (~40%). As shown in Table 4.2, $\overline{d_{S_{f_{ab}}}}/\overline{d_{S_{\psi_{ab}}}}$ is estimated to be less than one, which suggests that the slight deviation from neutrality is most likely not a result of selection.

In a previous study, Bustamante et al. (2002) estimated silent site evolution in functional genes (d_{S_f}) to be ~70% of the rate of evolution in pseudogenes (d_{S_ψ}). This is in stark contrast to the ~94% obtained in the present study. To reexamine the results obtained by Bustamante et al. (2002) I employed the same method and estimated the rate of synonymous substitution in human genes and pseudogenes using mouse or marmoset as an outgroup species (Table 4.3). When using mouse as the outgroup species $\overline{d_{S_f}}/\overline{d_{S_\psi}}$ is estimated to be between 0.80-0.83 and when using a more closely related outgroup such as marmoset it is estimated to be between 0.88-0.90 (Table 4.3). Similarly to the study by

Bustamante et al. (2002) when using distantly related species such as human and mouse $\overline{d_{S_f}}/\overline{d_{S_\psi}}$ is largely underestimated. Our observations are confirmed by simulations mimicking the formation of human processed pseudogenes after the human-mouse and human-marmoset split (Figure 4.2). As indicated in Figure 4.4, in all types of simulations $\overline{d_{S_f}}/\overline{d_{S_\psi}}$ is smaller than one. Furthermore, the ratio becomes even smaller when using a distantly related outgroup such as mouse.

To examine whether the larger difference observed, when using mouse as an outgroup, is a result of the change in GC content experienced by processed pseudogenes, I performed simulations in which the same equilibrium codon frequencies were used over the gene and pseudogene branches (note: These simulations were done using INDELIBLE with $\omega = 1$ for the pseudogene branch). The results showed no significant difference between $\overline{d_{S_f}}$ and $\overline{d_{S_\psi}}$ (results not shown). Bustamante et al. (2002) tried to remove the above effects by comparing genes and pseudogenes found in regions of similar GC content; however their approach may not have eliminated all possible effects.

Using both real and simulated sequences I thus uncovered the possibility that previous estimates of selection on synonymous sites is part driven by the change in GC content after pseudogene formation. Significant selection on synonymous sites, as indicated by codon bias, may only happen in species such as *Drosophila melanogaster, Saccharomyces cerevisiae, Arabidopsis thaliana, and Caenorhabditis elegans* (Hershberg and Petrov 2008) whose effective population sizes are in the order of several hundreds of thousands, or millions (Charlesworth 2009; Skelly et al. 2009; Cao et al. 2011). It could also be that in mammals, selection on synonymous sites is limited to a small number of proteins and is not common enough to be detected genome wide.

Chapter 5: Is there a relationship between generation time and selection in mammals?

Abstract

Theory predicts that selection against deleterious mutations should increase as effective population size increases. Using generation time as a proxy of effective population size, previous studies have observed that the ratio of the divergence of nonsynonymous to synonymous sites ($\omega = d_N/d_S$) increases as generation time increases. Using a large sample of orthologous protein-coding genes I reexamined the relation between generation time and selection in 13 high coverage mammals. In addition, I considered the effects of sequence quality and alignment quality on ω. My results indicate that mean selection within mammals can be separated into three groups with simians having the largest ω (~0.17), horse and elephant being intermediate (~0.14), and the rest of the mammals having the approximate same ratio (~0.11). Unlike previous studies, within the sample of 13 mammals, I did not observe a significant correlation between generation time and ω after accounting for phylogenetic dependence. Interestingly, within primates there is a significant decrease in the efficiency of selection in monkeys after their divergence from prosimians.

Introduction

One way to estimate selection on protein coding genes is by using the ratio of nonsynonymous to synonymous substitutions ($\omega = d_N/d_S$). The majority of protein coding genes are estimated to be under purifying selection ($\omega < 1$), with a very small proportion being under positive selection ($\omega > 1$) (Bakewell et al. 2007; Gibbs et al. 2007; Kosiol et al. 2008). The efficiency of purifying selection on deleterious mutations depends on the selection coefficient (s) and the effective population size (N_e) (Ohta and Kimura 1971). According to theory, deleterious mutations whose selection coefficients are lower than $1/N_e$ are subject to genetic drift, while deleterious mutations with selection coefficients are greater than $1/N_e$ are selected against (Kimura 1968a). Therefore, species with large effective population sizes should have a smaller proportion of effectively neutral mutations than species with small effective population sizes. Using generation time as a proxy to effective population size studies have shown that nonsynonymous sites are under stronger selection in species with shorter generation times or large effective population sizes than species with longer generation times or smaller effective population sizes (Ohta 1993; Eyre-Walker et al. 2002; Piganeau and Eyre-Walker 2009). These however, were limited to a small number of genes and comparisons were mainly made between species with different life history traits. Differences in d_N/d_S between species with different life history traits could result from differences in selection (s) and not effective population size (N_e).

With the recent sequencing of a large number of species, studies reinforced the positive correlation between generation time and ω, within mammals (Nikolaev et al. 2007; Popadin et al. 2013). Examining the results provided by Nikolaev et al. (2007) it seems that most variation in ω is found between primates and non-primates. As shown in Figure 5.1, species with ω ratios that are significantly above the mean are all primates with the exception of elephant. Here I wanted to reexamine whether the correlation is still significant after accounting for phylogenetic dependence (Felsenstein 1985).

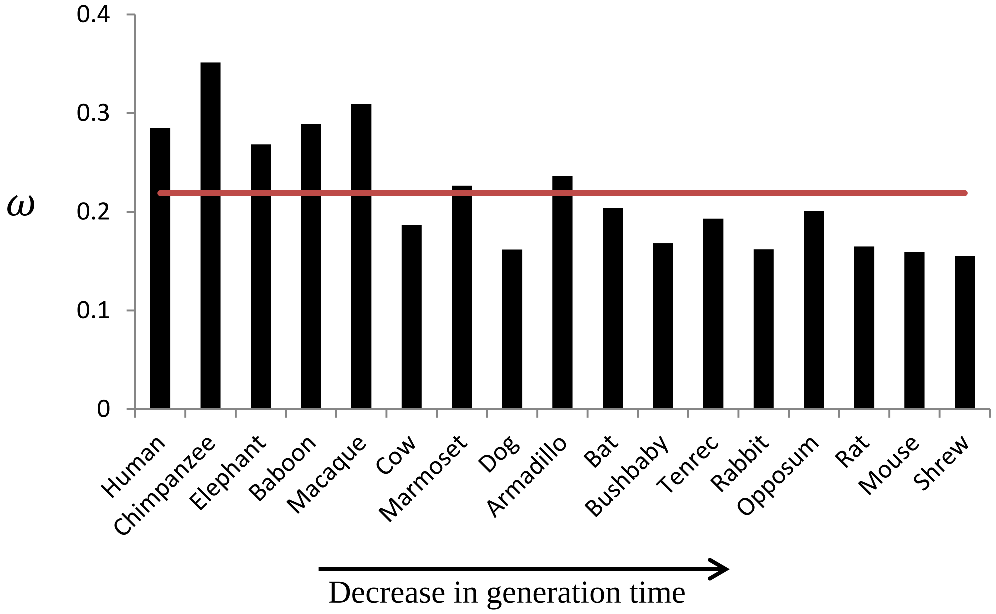

Figure 5.1. Estimates of ω from Nicolaev et al. (2007). Species are ranked according to the generation times provided in the study. The generation time for Opossum was obtained from Popadin et al. (2013). The red line depicts the mean ω value.

Another puzzling result is the fact that chimp shows an ω ratio that is 23% larger than human, even though chimps have a larger effective population size (Charlesworth 2009) and a similar generation time (Figure 5.1). A possible factor contributing to this discrepancy could be sequencing errors. Sequencing errors were previously shown to inflate d_N/d_S and resulted in wrongly predicted instances of positive selection (Schneider et al. 2009). Assuming that the rate of sequencing error is constant across species, their effect on ω should be more prominent in lineages with a small number of substitutions.

In this study, using a large sample of orthologous genes I reexamined the relation between generation time and selection and consider factors that can influence estimates of ω, particularly alignment quality and sequencing errors (Schneider et al. 2009; Fletcher and Yang 2010). Finally, given the recent sequencing of a number of prosimian genomes I also examined how the efficiency of selection varied within primates.

Materials and Methods

Assembling orthologous protein coding genes and estimating selection

Orthologous protein coding genes were downloaded using the OMA browser http://omabrowser.org/ (Schneider et al. 2007). One data set included 3,878 orthologous groups from 13 high coverage (>6X) mammalian species (human, chimp, orangutan, macaque, marmoset, mouse, rat, rabbit, cow, pig, horse, dog, elephant). This will be referred as "Mammalian" dataset. The other data set included 2,766 orthologs from 9 primate species (human, chimp, orangutan, macaque, marmoset, tarsier, bush baby, mouse lemur) in which tarsier and mouse lemur were of low coverage (≤2X). This will be referred as "Primate" dataset. Orthologs within each data set were aligned at the amino acid level using MAFFT (Katoh et al. 2002). The resulting alignments were duplicated, and for one of the sets badly aligned regions were masked using the alignment refinement mode of T-COFFEE (Notredame et al. 2000). Finally, amino acid alignments were used to build codon alignments. Codons corresponding to masked amino acids were not included.

For each of the codon alignments, selection was estimated using the free ratio model in CODEML (Yang 1997) and the two phylogenies depicted in Figure 5.2. The phylogeny was trimmed according to the species being examined. After obtaining lineage specific estimates of selection, I only kept groups of othologs that met the following criteria: (1) all species had an $\omega < 1$; (2) the number of synonymous sites was ≥100; (3) d_S values were $0 < d_S < 0.5$. The condition $\omega < 1$ was used to remove genes that may evolve under positive selection. Using sequences with at least 100 synonymous sites reduced errors due to small sample size. The condition of $d_S > 0$ is necessary to estimate selection, and finally the condition $d_S < 0.5$ limited inflated estimates. After the above filtering, I was left with estimates from 1,683 orthologs for the high coverage "Mammalian" data set, and 1,417 for the "Primate" dataset.

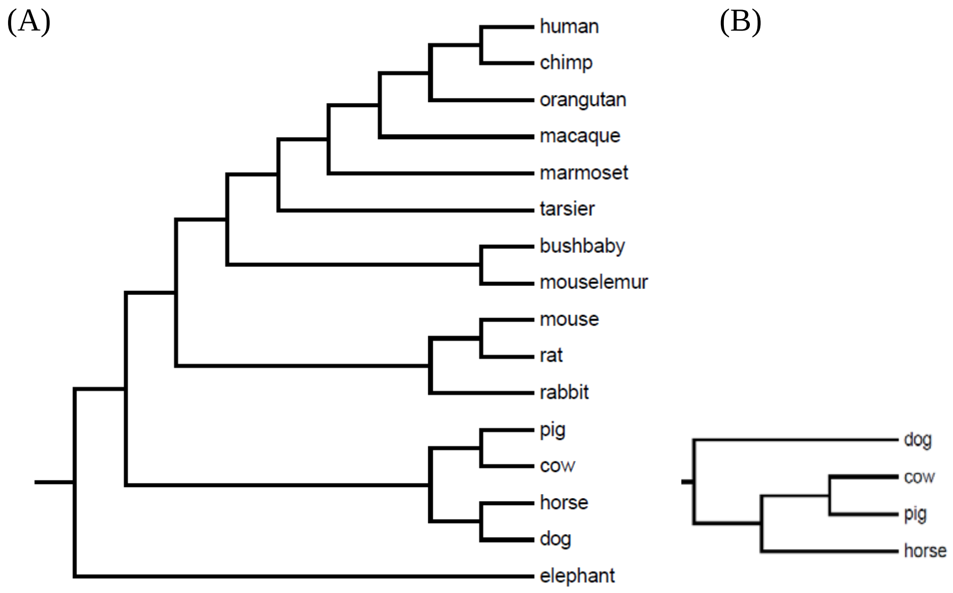

Figure 5.2: The alternative phylogenies for the species under study. In (B) are the differences from (A) (Prasad et al. 2008).

A brief explanation of the OMA algorithm

Because our sample of orthologous genes seems small, we would like to emphasize that the quantity is compensated for quality, i.e., the degree of confidence in the data. This has to do with the manner in which the OMA algorithm works (Schneider et al. 2007). The steps used by the OMA algorithm to detect orthologs are as follows:

(1) All-against-all: Sequences belonging to complete genomes of eukaryotes are downloaded from Ensembl. An all against all comparison is made using local dynamic programming (Smith-Waterman). To remove spurious hits they only keep alignments that cover 69% of the longest sequence.

(2) For each pair of genomes the algorithm detects stable pairs. Stables pairs are pairs of sequences that have no significantly closer homolog than each other.

(3) The next step involves detecting paralogs. For each stable pair identified above, a third genome is used to detect paralogs. In Figure 5.3, a paralogous relation is detected between genes X_1 and Y_1 using an eye witness genome Z. This step is done for every eye witness genome that carries orthologs to genes X_1 and Y_1, resulting in the removal of paralogs and a final dataset in which sequences are 1:1 orthologs.

(4) The final step is the formation of OMA groups. In an OMA group, every protein is orthologous to every other protein in the group. Consequently an OMA group has at most one sequence from each genome

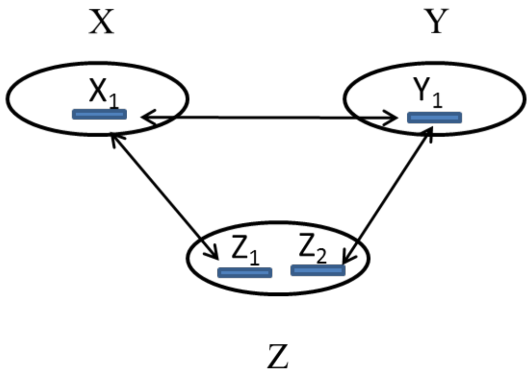

Figure 5.3: Given a stable gene pair X_1 and Y_1, a third genome Z is used to test whether X_1 and Y_1 are orthologs. In this case X_1 and Y_1 show significant similarity to two different genes in genome Z, therefore, they are considered paralogs.

Assessing Sequence Quality

Sequence quality was estimated two ways. One way used whole genome shotgun reads to estimate the coverage for each CDS downloaded from the OMA browser. The method employed was a slightly modified version of the method introduced by Schneider et al. (2009). Specifically, for each species I downloaded its trace data (Trace archive, http://www.ncbi.nlm.nih.gov/Traces/home/). Each species set of CDSs was BLASTed against its corresponding trace archive. Only hits with a 98% percent identity and an e-value of $<10^{-10}$ were considered, in order to reduce random hits or hits to reads belonging to duplicated genes. Sequence coverage was estimated as the number of hits overlapping a position in the CDS. Sequences were considered as high coverage if ≥95% of their nucleotides were read at least three or more times.

The second method employed quality scores for each base. These came in two forms: (1) raw quality scores and (2) PHRED scores. Both of these scores estimate the base calling accuracy for a particular base. According to the UCSC genome browser a raw quality score between 45-97 is considered high quality (http://genome.ucsc.edu/FAQ/FAQformat.html). A PHRED score ranges from 10-50 with a score of 50 representing a 99.999% base calling accuracy (Lander et al. 2001). Both types of quality scores were downloaded from the UCSC genome browser (Karolchik et al. 2014) for all primates except bush baby. Quality scores for bush baby were not available but the new version of the genome is considered high coverage (137 X) (Di Palma et al. 2011). In genomes where the quality of base calling was based on a PHRED score I defined high quality sequences as ones in which ≥ 95% of their bases had a score of 40 and above (99.99% base calling accuracy). When using a raw quality score I defined high quality sequences as ones in which ≥ 95% of their bases had the highest score found within the genome. These ranged from 45-97 depending on the species.

Human and mouse genomes were not included in the analysis because their genomes were assembled using bacterial artificial chromosome (BAC) sequencing (Lander et al. 2001; Church et al. 2009), and therefore are qualitatively superior to the draft assemblies of other species based on whole genome shotgun (WGS) sequencing.

Estimates of generation time

The two most common definitions of generation length (time) are: (a) the average age of parents in the current cohort and (b) the age at which half of total reproductive output is achieved by an individual (Pacifici et al. 2013). In the present study I built two sets of two sets of generation times which I name “GT_1” and “GT_2”. GT_1 was built using generation times obtained from the study by Pacifici et.al (2013). In GT_1 the generation time for human was obtained by the study by Langergraber et.al (2012). Generation time in these studies was regarded as the average age of parents in the current cohort. Because the generation time of dog was not available, we used the estimated generation time of the gray wolf. The dataset GT_2 was built using years to female sexual maturity (Jones et al. 2009) as a proxy to generation time. GT_{orig} are the generation times used by Nikolaev et.al (2007). All data is depicted in Table 5.1.

Table 5.1: Estimates of generation time obtained for 20 mammals.

	GT_1	GT_2	GT_{orig}
human	29	15	20
chimp	20	11	15
orangutan	25	9	NA
macaque	14	3	5
baboon	16	7	6
marmoset	6	1	2
tarsier	6	1	NA
mouse lemur	6	1	NA
bush baby	6	2	1
rabbit	4	0.5	0.5
mouse	1.6	0.2	0.1
rat	1.7	0.2	0.1
horse	8	2	NA
dog	8	2	2
cow	9	2	3
pig	7	1	NA
tenrec	5	1	0.5
elephant	25	14	12
armadillo	5	1	1
opossum	2	0.4	0.3

Testing for phylogenetic noindependence

To remove the effects of phylogenetic dependence I used the independent contrasts test (Felsenstein 1985) implemented in COMPARE (Martins 2004). The phylogenies used for the tests, were downloaded from the UCSC genome browser (http://genome-source.cse.ucsc.edu/gitweb/?p=kent.git;a=tree;f=src/hg/utils/phyloTrees). Branch lengths were estimated using whole genome alignments.

Estimating mean selection on nonsynonymous sites

Mean selection on nonsynonymous sites ($\bar{\omega}$) was expressed as $\overline{d_{N_w}}/\overline{d_{S_w}}$, where $\overline{d_{N_w}}$ and $\overline{d_{S_w}}$ represent the weighted averages of nonsynonymous and synonymous substitutions, respectively. Estimates were weighted according to the number of nonsynonymous and synonymous sites respectively. The use of weighted averages allows us to give more weight to longer sequences whose estimates of d_N and d_S would be subjected to lesser degrees of random error. To generate the 95% confidence intervals for $\bar{\omega}$, I used 10,000 bootstrap samples with replacement. For each sample, I computed $\overline{d_{N_w}}/\overline{d_{S_w}}$ and estimated the mean and 95% confidence intervals using the bootstrap distribution.

Results

Comparing estimates of selection between low-quality and high-quality sequences

Rates of nonsynonymous and synonymous substitutions that were estimated from the 1,683 orthologs in mammals were separated into ones corresponding to high coverage coding sequences and ones corresponding to low coverage coding sequences. In the case of human and mouse all sequences were considered to be high-coverage. The $\bar{\omega}$ was estimated for each category (Figure 5.4 A).

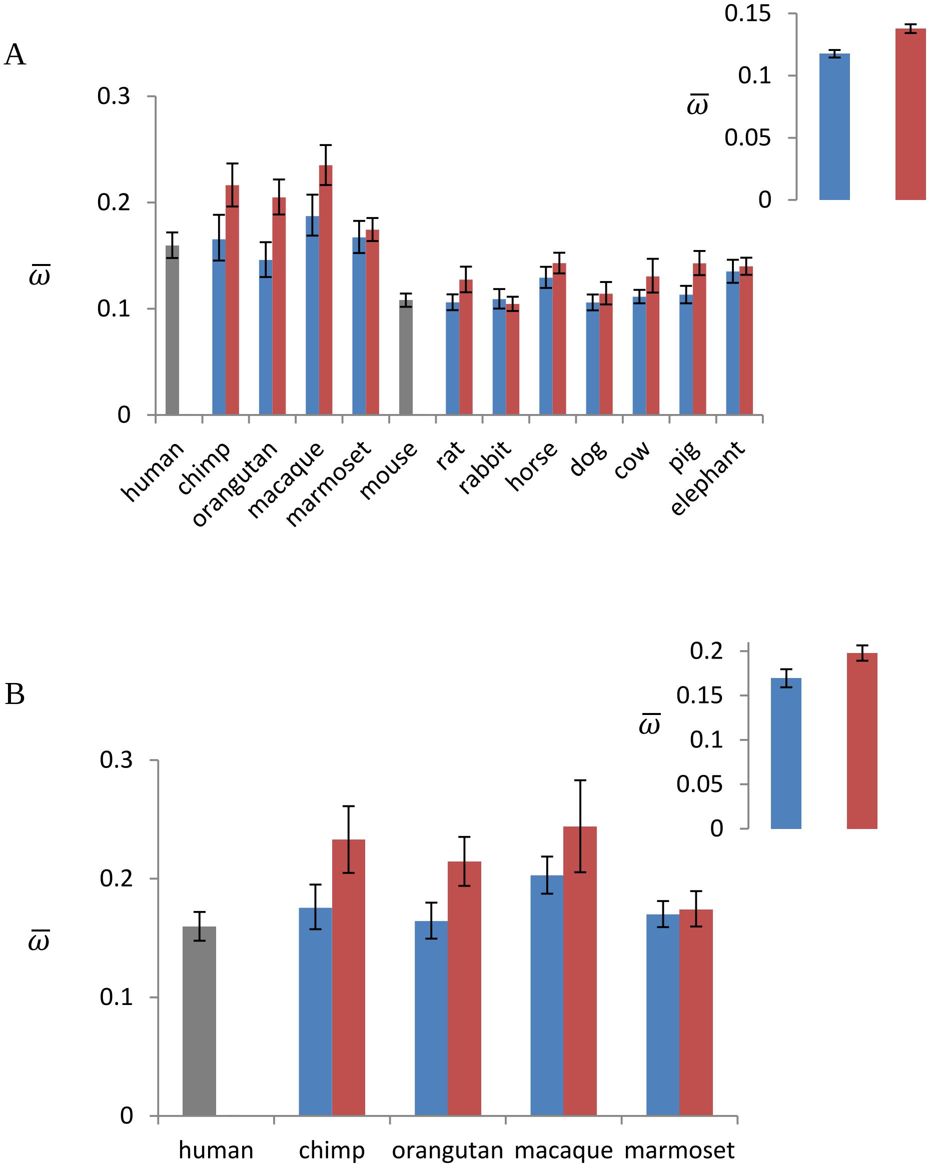

Figure 5.4: Mean selection ($\overline{\omega}$) estimated by grouping ω values from high quality (blue) and low quality (red) orthologous sequences in mammals ("Mammalian" dataset). (A) Quality estimated using WGS reads for all mammals except human and mouse (B) For primates only, quality was also estimated using raw quality scores. All species data are plotted on the right hand corner. Error bars represent 95% confidence intervals. Gray bars represent high quality sequences.

Differences between low and high quality data were more prominent within primates. For example, in human, chimp, and orangutan, $\bar{\omega}$ values from low coverage sequences were on average ~32% larger than in high coverage sequences. In Primates, I reexamined the results using raw quality scores (Figure 5.4 B). The same trend was observed using both types of methods.

Finally, I looked at the effects of quality on ω in set of primate orthologs including prosimians. A significant difference between low and high quality sequences was only observed in great apes (Figure 5.5).

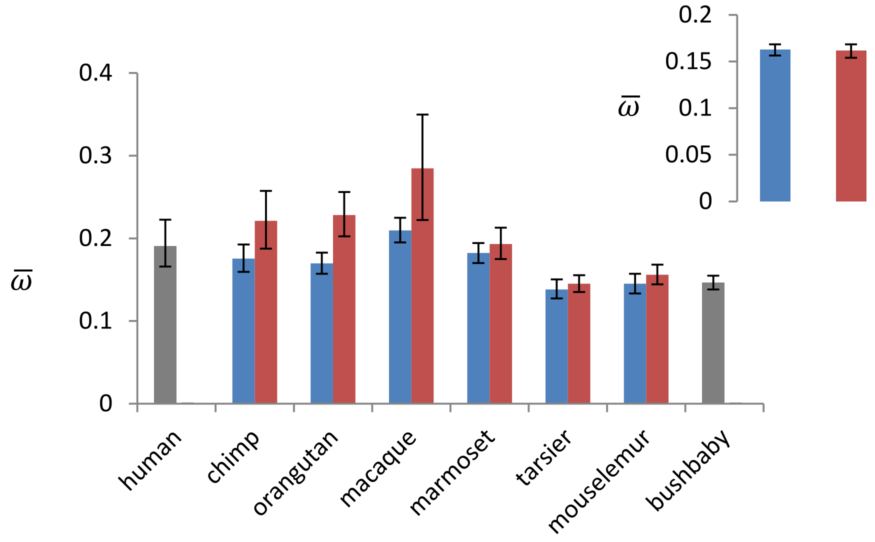

Figure 5.5: Comparison of mean selection estimated using high quality (blue) and low quality (red) orthologous sequences in primates (“Primate” dataset). Quality was estimated using raw quality scores in simians and PHRED scores in prosimians. Error bars represent 95% confidence intervals.

Alignment quality and estimating selection

To examine whether removing badly aligned regions changes estimates of mean selection I used a common set of 1,632 orthologs and tested whether selection differed with or without using alignment refinement. To remove effects of sequence quality I only used high quality data in each species. As shown in Figure 5.6 mean selection does not significantly change when considering alignment quality, however, there is a trend where $\overline{\omega}$ becomes smaller after alignment refinement. This trend causes a significant difference of ~10% when considering all species (Figure 5.6, right corner).

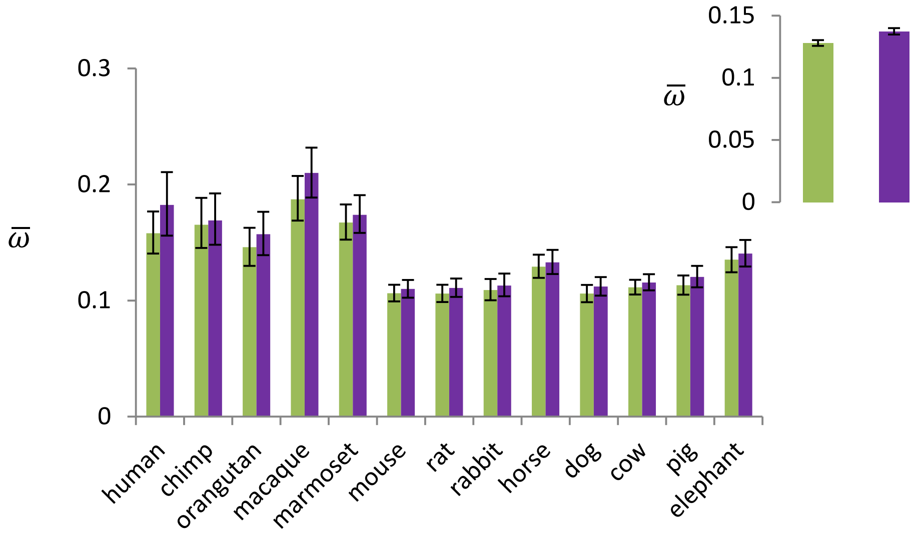

Figure 5.6: Mean selection estimated using all codons (purple) versus codons with a high alignment score (green). Error bars represent 95% confidence intervals.

Comparing selection within mammals

Table 5.2A lists mean purifying selection ($\overline{\omega}$) values in thirteen mammals. Estimates were based on high quality sequences. Results in Table 5.2A can be separated into three groups depending on similarity in selection. Members of each group do not show any significant difference in selection, except in the case of macaque. The average $\bar{\omega}$ ratios within these groups are approximately 0.17, 0.13, and 0.11, respectively. When examining the correlation between generation time and selection using data from the present study and Nikolaev et.al's (2007) (Table 5.2B), I observe a significant positive correlation (Figure 5.7). Correlation remains significant even when using different sets of generation times (Table 5.3).

Table 5.2: Estimates of selection used in correlation analysis. (A) Selection ($\bar{\omega}$) in 13 mammals (Materials and Methods: "Mammalian dataset"). (B) d_N/d_S in 16 mammals. Values obtained from the study by Nikolaev et.al (2007).

(A)

	$\bar{\omega}$	CI^{95}
human	0.160	(0.141, 0.176)
chimp	0.165	(0.145, 0.189)
orangutan	0.146	(0.131, 0.163)
macaque	0.187	(0.169, 0.207)
marmoset	0.167	(0.152, 0.183)
horse	0.129	(0.120, 0.140)
elephant	0.135	(0.124, 0.146)
dog	0.106	(0.099, 0.113)
cow	0.111	(0.105, 0.118)
pig	0.113	(0.105, 0.122)
rabbit	0.109	(0.100, 0.119)
mouse	0.106	(0.099, 0.114)
rat	0.106	(0.099, 0.114)

(B)

	ω
human	0.29
chimp	0.35
baboon	0.29
macaque	0.31
marmoset	0.23
bushbaby	0.17
mouse	0.16
rat	0.16
rabbit	0.16
cow	0.19
dog	0.16
bat	0.20
shrew	0.16
armadillo	0.24
tenrec	0.19
elephant	0.27
monodelphis	0.20

(A) (B)

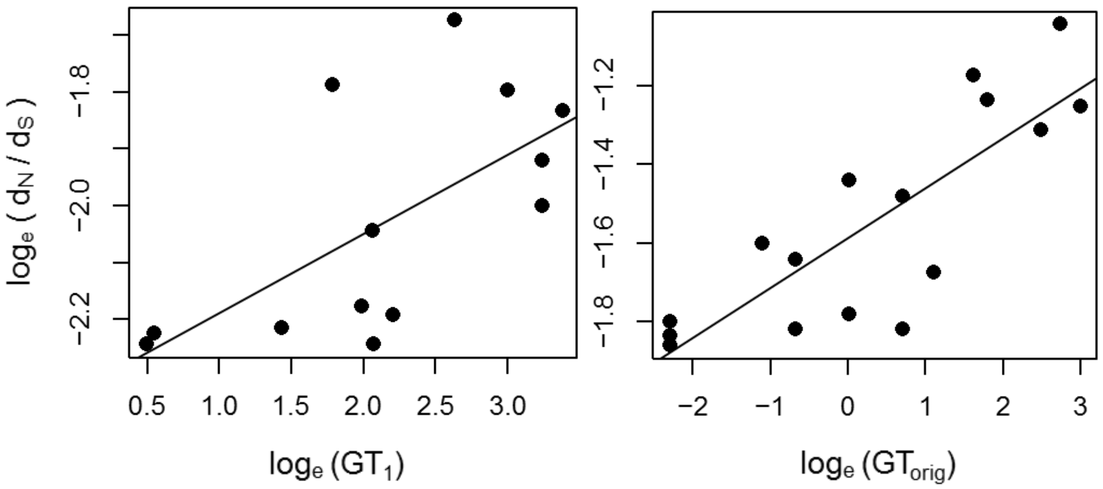

Figure 5.7: Linear regressions of generation time and selection. (A) Data from present study. Based on 13 high coverage mammals (Table 5.1A) the Pearson correlation was significant (r_p = 0.64, p-value = 0.0173). (B) Data on selection and generation were obtained from the study by Nikolaev et. al (2007) (r_p = 0.82, p-value < 0.001). These were based on 16 mammals except bat. Bat was excluded because its topology in the mammalian phylogeny is intensely debated (Tsagkogeorga et al. 2013).

After applying Felsenstein's independent contrasts test (1985) the correlations between generation time and selection identified by the present study became insignificant (Table 5.3). After applying the phylogenetic contrast test to Nikolaev et al's data the correlations were much larger than the present study but still insignificant (Table 5.3). Based on a two tailed t-test with 14 degrees of freedom, the correlations r = 0.38 and r = 0.19 gives a test statistic of 1.8 and 0.79.

Table 5.3: Correlations between selection and generation time (GT_1, GT_2, GT_{orig} are different sets of generation times)

	Present study (N=13)		Nikolaev et.al (N=16)		
	GT_1	GT_2	GT_{orig}	GT_1	GT_2
r_p	0.64*	0.64*	0.82*	0.79*	0.79*
r_s	0.56*	0.65*	0.79*	0.75*	0.70*
r_{pic}	0.00	0.05	0.38	0.19	0.19

$r_{p:}$ pearson; $r_{p:}$ spearman; $r_{pic:}$ phylogenetically independent contrast (* indicates significance)

When examining selection in primates (Table 5.4), there is a significant difference in selection between simians and prosimians. In addition, within simians macaque shows a significant difference.

Table 5.4: Selection estimated in great apes, monkeys, and prosimians (Materials and Methods: "Primate dataset").

	$\bar{\omega}$	CI^{95}
human	0.191	(0.166,0.223)
chimp	0.176	(0.160,0.192)
orangutan	0.17	(0.157,0.183)
macaque	0.21	(0.195,0.225)
marmoset	0.182	(0.170,0.194)
tarsier	0.138	(0.127,0.150)
bush baby	0.146	(0.138,0.155)
mouse lemur	0.135	(0.126,0.145)

Discussion

In this study, I compared selection across the mammalian phylogeny. Selection was estimated over a group of thirteen high coverage mammals. As shown in Figures 5.4 and 5.5 sequencing errors tend to inflate mean purifying selection. This is more prominent in the case of primates. The larger effect in primates is most likely caused by sequencing errors having a greater impact on branches with a small number of changes. In contrast to sequencing quality, alignment quality did not have a significant impact on average selection but caused a trend of lower $\overline{\omega}$ values (Figure 5.6).

After accounting for sequence quality, I examined how selection varied across mammalian species. As shown in Table 5.2A, in terms of average selection, the species examined could be separated into three groups. As in previous studies, primates showed the least efficiency against deleterious mutations (Nikolaev et al. 2007; Popadin et al. 2013) with ~17% being effectively neutral. Species with the highest efficiency against deleterious mutations ($\overline{\omega} \approx 0.11$) were from different mammalian orders including Glires, Carnivora and Artiodactyla. In accordance to the negative correlation between generation time and effective population size primates show the highest $\overline{\omega}$ (Table 5.1A).

Previous studies in mammals showed a positive linear relation between generation time and selection (Nikolaev et al. 2007; Popadin et al. 2013). In the present study I reexamined the relation between generation time and selection, using data from the study by Nikolaev et al. (2007) and the 13 high coverage mammals studied in the present study.

As seen in Figure 5.7 both data sets show a significant correlation. However, after applying Felsenstein’s independence contrast test, correlations become insignificant (Table 5.3). The results remained the same even when using different sets of generation times. Furthermore, results remained insignificant, when omitting alignment refinement or when using all sequences (Data not shown).

To examine whether any unresolved parts of the mammalian phylogeny had an impact on the results I examined selection using a phylogeny that supported the earlier divergence of Carnivora from Perissodactyla and Artiodactyla (Prasad et al. 2008). The results indicated no significant difference (Data not shown).

The lack of any correlation between generation time and selection may result from the limited number of data from non-primate species with long generation times. With the sequencing of more mammalian genomes the correlation may gain support. An alternative explanation maybe that generation time is not a good proxy of effective population size. In the future, I plan to examine whether there is a relation between selection and long term effective population size.

Interestingly, within simians macaque showed a consistently larger ω ratio, even when accounting for sequence quality (Table 5.2A). Although this can be due to differences in selection it could also be due to errors during genome assembly. An unpublished study that sequenced a new version of the rhesus genome (Zimin et al. 2013) found a 97%

percent identity with human orthologs while the older version used in the present study showed an 88% similarity (Gibbs et al. 2007).

Given that most simians did not show any significant differences in selection I tested for differences in selection between simians and prosimians. As shown in Table 5.4 there is a significant difference in selection between prosimians and simians. Specifically in simians the average $\bar{\omega}$ ratio excluding macaque is ~30% larger than prosimians. Assuming there are no significant effects caused by sequencing errors or wrongly annotated orthologs this may signify a significant reduction in the historical effective population size of simians.

Chapter 6: Summary

The efficiency of selection on deleterious and advantageous mutations is dependent on the product of the effective population size and selection coefficient (Wright 1931; Kimura 1968a). Given the very low mutation rates and effective population sizes in most animals selection for mutational robustness is expected to be effectively neutral. Contrary, to previous claims of selection for mutational robustness in mammalian pre-miRNAs, (Borenstein and Ruppin 2006), I show that mutational robustness in *Drosophila* pre-miRNAs evolved neutrally. In addition, I show that mutational robustness did not evolve as a by-product of selection for robustness against thermal fluctuations (Ancel and Fontana 2000; Szollosi and Derenyi 2009) but most likely as a byproduct of selection to preserve structure and function.

In Chapter 3 I identify orthologous processed pseduogenes and use them as a neutral model to test for selection on synonymous sites. Contrary to previous studies that estimated ≥25% of synonymous sites being under selection (Ophir et al. 1999; Bustamante et al. 2002; Hellmann et al. 2003; Eory et al. 2010) I identify a percentage that is closer to ≤8%. In addition, using sequence simulations I show that the deviations from neutrality in the current study and the study by Bustamante et al. (2002) could result because processed pseudogenes experience a shift in GC after they are formed.

In Chapter 5 using a 13 high coverage genomes I examined the effects of sequencing and alignment errors when estimating purifying selection on nonsynonymous sites. Although alignment errors do not have any significant effects, sequencing errors tend to inflate d_N/d_Sratios. After accounting for sequencing errors I re-examined the negative correlation between generation time and effective population size using the rate of fixation of deleterious alleles at nonsynonymous sites. As expected species with smaller generation times or large effective population sizes tend to be more efficient in purging deleterious mutations than species with longer generation times or small effective population sizes. However, there is no significant evidence of a linear relation between d_N/d_S and generation time. In addition, I detect a significant shift in purifying selection after prosimians diverged from new world monkeys. With the advent of more high quality mammalian genomes, re-examining the positive correlation between generation time and selection should become more powerful and accurate.

References

Abrusan G, Giordano J, Warburton PE. 2008. Analysis of transposon interruptions suggests selection for L1 elements on the X chromosome. PLoS genetics 4:e1000172.

Akashi H, Eyre-Walker A. 1998. Translational selection and molecular evolution. Curr Opin Genet Dev 8:688-693.

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

Ancel LW, Fontana W. 2000. Plasticity, evolvability, and modularity in RNA. J Exp Zool 288:242-283.

Anderson T, Darling D. 1952. Asymptotic theory of certain “goodness of fit” criteria based on stochastic processes. The Annals of Mathematical Statistics.

Azevedo RB, Lohaus R, Srinivasan S, Dang KK, Burch CL. 2006. Sexual reproduction selects for robustness and negative epistasis in artificial gene networks. Nature 440:87-90.

Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H. 2006. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol 2:2006 0008.

Bailey JA, Eichler EE. 2006. Primate segmental duplications: crucibles of evolution, diversity and disease. Nat Rev Genet 7:552-564.

Bakewell MA, Shi P, Zhang J. 2007. More genes underwent positive selection in chimpanzee evolution than in human evolution. Proc Natl Acad Sci U S A 104:7489-7494.

Bartel DP. 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116:281-297.

Blanchette M, Diallo AB, Green ED, Miller W, Haussler D. 2008. Computational reconstruction of ancestral DNA sequences. Methods Mol Biol 422:171-184.

Bloom JD, Silberg JJ, Wilke CO, Drummond DA, Adami C, Arnold FH. 2005. Thermodynamic prediction of protein neutrality. Proc Natl Acad Sci U S A 102:606-611.

Bonnet E, Wuyts J, Rouze P, Van de Peer Y. 2004. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. Bioinformatics 20:2911-2917.

Borenstein E, Ruppin E. 2006. Direct evolution of genetic robustness in microRNA. Proc Natl Acad Sci U S A 103:6593-6598.

Buratti E, Baralle FE. 2004. Influence of RNA secondary structure on the pre-mRNA splicing process. Mol Cell Biol 24:10505-10514.

Bustamante CD, Nielsen R, Hartl DL. 2002. A maximum likelihood method for analyzing pseudogene evolution: implications for silent site evolution in humans and rodents. Mol Biol Evol 19:110-117.

Caceres EF, Hurst LD. 2013. The evolution, impact and properties of exonic splice enhancers. Genome Biol 14:R143.

Cao J, Schneeberger K, Ossowski S, Gunther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C, et al. 2011. Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat Genet 43:956-963.

Carlini DB, Genut JE. 2006. Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers. J Mol Evol 62:89-98.

Cartwright RA. 2005. DNA assembly with gaps (Dawg): simulating sequence evolution. Bioinformatics 21 Suppl 3:iii31-38.

Chamary JV, Hurst LD. 2005. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. Genome Biol 6:R75.

Chamary JV, Parmley JL, Hurst LD. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet 7:98-108.

Charchar FJ, Svartman M, El-Mogharbel N, Ventura M, Kirby P, Matarazzo MR, Ciccodicola A, Rocchi M, D'Esposito M, Graves JA. 2003. Complex events in the evolution of the human pseudoautosomal region 2 (PAR2). Genome Res 13:281-286.

Charlesworth B. 2009. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. Nat Rev Genet 10:195-205.

Church DM, Goodstadt L, Hillier LW, Zody MC, Goldstein S, She X, Bult CJ, Agarwala R, Cherry JL, DiCuccio M, et al. 2009. Lineage-specific biology revealed by a finished genome assembly of the mouse. PLoS Biol 7:e1000112.

Churkin A, Cohen M, Shemer-Avni Y, Barash D. 2010. Bioinformatic analysis of the neutrality of RNA secondary structure elements across genotypes reveals evidence for direct evolution of genetic robustness in HCV. Journal of bioinformatics and computational biology 8:1013-1026.

Comeron JM. 2006. Weak selection and recent mutational changes influence polymorphic synonymous mutations in humans. Proc Natl Acad Sci U S A 103:6940-6945.

Datta A, Jinks-Robertson S. 1995. Association of increased spontaneous mutation rates with high levels of transcription in yeast. Science 268:1616-1619.

de Visser JA, Hermisson J, Wagner GP, Ancel Meyers L, Bagheri-Chaichian H, Blanchard JL, Chao L, Cheverud JM, Elena SF, Fontana W, et al. 2003. Perspective: Evolution and detection of genetic robustness. Evolution 57:1959-1972.

Dewey CN, Rogozin IB, Koonin EV. 2006. Compensatory relationship between splice sites and exonic splicing signals depending on the length of vertebrate introns. BMC Genomics 7:311.

Di Palma F, Johnson J, Lander ES, Lindblad-Toh K, Jaffe DB, Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, et al. 2011. Version 3 of the genome sequence of Otolemur garnettii. Unpublished.

Diallo AB, Makarenkov V, Blanchette M. 2010. Ancestors 1.0: a web server for ancestral sequence reconstruction. Bioinformatics 26:130-131.

Doherty A, McInerney JO. 2013. Translational selection frequently overcomes genetic drift in shaping synonymous codon usage patterns in vertebrates. Mol Biol Evol 30:2263-2267.

Drake JW, Charlesworth B, Charlesworth D, Crow JF. 1998. Rates of spontaneous mutation. Genetics 148:1667-1686.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134:341-352.

Duret L. 2002. Evolution of synonymous codon usage in metazoans. Curr Opin Genet Dev 12:640-649.

Ehrenreich IM, Purugganan MD. 2008. Sequence variation of MicroRNAs and their binding sites in Arabidopsis. Plant Physiol 146:1974-1982.

Emerson JJ, Kaessmann H, Betran E, Long M. 2004. Extensive gene traffic on the mammalian X chromosome. Science 303:537-540.

Eory L, Halligan DL, Keightley PD. 2010. Distributions of selectively constrained sites and deleterious mutation rates in the hominid and murid genomes. Mol Biol Evol 27:177-192.

Esnault C, Maestre J, Heidmann T. 2000. Human LINE retrotransposons generate processed pseudogenes. Nat Genet 24:363-367.

Eyre-Walker A, Keightley PD, Smith NG, Gaffney D. 2002. Quantifying the slightly deleterious mutation model of molecular evolution. Mol Biol Evol 19:2142-2149.

Fairbrother WG, Holste D, Burge CB, Sharp PA. 2004. Single nucleotide polymorphism-based validation of exonic splicing enhancers. PLoS Biol 2:E268.

Felsenstein J. 1989. PHYLIP-Phylogeny Inference Package. Cladistics:164-166.

Felsenstein J. 1985. Phylogenies and the comparative method. Amer Nat 125:1-15.

Flatt T. 2005. The evolutionary genetics of canalization. Q Rev Biol 80:287-316.

Fletcher W, Yang Z. 2010. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. Mol Biol Evol 27:2257-2267.

Fletcher W, Yang Z. 2009. INDELible: a flexible simulator of biological sequence evolution. Mol Biol Evol 26:1879-1888.

Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, et al. 2013. Ensembl 2013. Nucleic Acids Res 41:D48-55.

Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, et al. 2012. Ensembl 2012. Nucleic Acids Res 40:D84-90.

Forster R, Adami C, Wilke CO. 2006. Selection for mutational robustness in finite populations. J Theor Biol 243:181-190.

Frith MC, Wilming LG, Forrest A, Kawaji H, Tan SL, Wahlestedt C, Bajic VB, Kai C, Kawai J, Carninci P, et al. 2006. Pseudo-messenger RNA: phantoms of the transcriptome. PLoS Genet 2:e23.

Gardner A, Kalinka AT. 2006. Recombination and the evolution of mutational robustness. J Theor Biol 241:707-715.

Gibbs RA, Rogers J, Katze MG, Bumgarner R, Weinstock GM, Mardis ER, Remington KA, Strausberg RL, Venter JC, Wilson RK, et al. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. Science 316:222-234.

Gibson G, Wagner G. 2000. Canalization in evolutionary genetics: a stabilizing theory? Bioessays 22:372-380.

Gould S. 1989. Wonderful life: the Burgess Shale and the nature of history. New York: W. W. Norton & Company.

Graur D. 2014. Molecular and Genome Evolution. Sunderland, MA: Sinauer.

Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. 2008. miRBase: tools for microRNA genomics. Nucleic Acids Res 36:D154-158.

Gu W, Wang X, Zhai C, Xie X, Zhou T. 2012. Selection on synonymous sites for increased accessibility around miRNA binding sites in plants. Mol Biol Evol 29:3037-3044.

Guo HH, Choe J, Loeb LA. 2004. Protein tolerance to random amino acid change. Proc Natl Acad Sci U S A 101:9205-9210.

Han J, Lee Y, Yeom KH, Nam JW, Heo I, Rhee JK, Sohn SY, Cho Y, Zhang BT, Kim VN. 2006. Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. Cell 125:887-901.

Harrison PM, Zheng D, Zhang Z, Carriero N, Gerstein M. 2005. Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability. Nucleic Acids Res 33:2374-2383.

Hellmann I, Zollner S, Enard W, Ebersberger I, Nickel B, Paabo S. 2003. Selection on human genes as revealed by comparisons to chimpanzee cDNA. Genome Res 13:831-837.

Hershberg R, Petrov DA. 2008. Selection on codon bias. Annual review of genetics 42:287-299.

Hietpas RT, Jensen JD, Bolon DN. 2011. Experimental illumination of a fitness landscape. Proc Natl Acad Sci U S A 108:7896-7901.

Hofacker I, Fontana W, Stadler P, Bonhoeffer S, Tacker M, Schuster P. 1994. Fast folding and comparison of RNA secondary structures. Monatshefte für Chemie 125.

Hurst LD. 2006. Preliminary assessment of the impact of microRNA-mediated regulation on coding sequence evolution in mammals. J Mol Evol 63:174-182.

Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2:13-34.

Jones EK, Bielby J, Cardillo M, Susanne FA, O'Dell J, Orme CDL, Safi K, Sechrest W, Boakes HE, Carbone C, et al. 2009. PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals Ecology 90:2648-2648.

Kanaya S, Yamada Y, Kinouchi M, Kudo Y, Ikemura T. 2001. Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. J Mol Biol 53:290-298.

Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, et al. 2014. The UCSC Genome Browser database: 2014 update. Nucleic Acids Res 42:D764-770.

Karro JE, Yan Y, Zheng D, Zhang Z, Carriero N, Cayting P, Harrrison P, Gerstein M. 2007. Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. Nucleic Acids Res 35:D55-60.

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.

Katoh K, Toh H. 2008. Recent developments in the MAFFT multiple sequence alignment program. Brief Bioinform 9:286-298.

Khurana E, Lam HY, Cheng C, Carriero N, Cayting P, Gerstein MB. 2010. Segmental duplications in the human genome reveal details of pseudogene formation. Nucleic Acids Res 38:6997-7007.

Kimura M. 1968a. Evolutionary rate at the molecular level. Nature 217:624-626.

Kimura M. 1968b. Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. Genet Res 11:247-269.

Kimura M. 1967. On the evolutionary adjustment of spontaneous mutation rates. Genet Res 9:23-34.

Kimura M. 1962. On the probability of fixation of mutant genes in a population. Genetics 47:713-719.

Kitano H. 2004a. Biological robustness. Nat Rev Genet 5:826-837.

Kitano H. 2004b. Cancer as a robust system: implications for anticancer therapy. Nat Rev Cancer 4:227-235.

Kiyohara M, Koyanagi T, Matsui H, Yamamoto K, Take H, Katsuyama Y, Tsuji A, Miyamae H, Kondo T, Nakamura S, et al. 2012. Changes in microbiota population during fermentation of narezushi as revealed by pyrosequencing analysis. Bioscience, biotechnology, and biochemistry 76:48-52.

Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A. 2008. Patterns of positive selection in six Mammalian genomes. PLoS Genet 4:e1000144.

Kozak M. 1999. Initiation of translation in prokaryotes and eukaryotes. Gene 234:187-208.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860-921.

Lawrie DS, Messer PW, Hershberg R, Petrov DA. 2013. Strong Purifying Selection at Synonymous Sites in D. melanogaster. PLoS Genet 9:e1003527.

Li W-H. 1997. Molecular evolution. Sunderland, Mass.: Sinauer Associates.

Li WH. 1978. Maintenance of Genetic Variability under the Joint Effect of Mutation, Selection and Random Drift. Genetics 90:349-382.

Liang H, Li WH. 2009. Lowly expressed human microRNA genes evolve rapidly. Mol Biol Evol 26:1195-1198.

Lind PA, Berg OG, Andersson DI. 2010. Mutational robustness of ribosomal protein genes. Science 330:825-827.

Lynch M, Conery JS. 2003. The origins of genome complexity. Science 302:1401-1404.

Marsaglia G, Marsaglia J. 2004. Evaluating the Anderson-Darling distribution. Journal Statistical Software 9.

Martins EP. 2004. COMPARE, version 4.6b. Computer programs for the statistical analysis of comparative data. Distributed by the author at http://compare.bio.indiana.edu/. In. Bloomington IN: Department of Biology, Indiana University.

Meiklejohn C, Hartl D. 2002. A single mode of canalization. Trends Ecol Evolut 17:468-473.

Miller JH. 1979. Genetic studies of the lac repressor. XI. On aspects of lac repressor structure suggested by genetic experiments. J Mol Biol 131:249-258.

Misevic D, Ofria C, Lenski RE. 2006. Sexual reproduction reshapes the genetic architecture of digital organisms. Proc Biol Sci 273:457-464.

Montville R, Froissart R, Remold SK, Tenaillon O, Turner PE. 2005. Evolution of mutational robustness in an RNA virus. PLoS Biol 3:e381.

Nachman MW, Crowell SL. 2000. Estimate of the mutation rate per nucleotide in humans. Genetics 156:297-304.

Nikolaev SI, Montoya-Burgos JI, Popadin K, Parand L, Margulies EH, Antonarakis SE. 2007. Life-history traits drive the evolutionary rates of mammalian coding and noncoding genomic elements. Proc Natl Acad Sci U S A 104:20443-20448.

Ninio J. 1991. Connections between translation, transcription and replication error-rates. Biochimie 73:1517-1523.

Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. Journal of molecular biology 302:205-217.

Nozawa M, Miura S, Nei M. 2010. Origins and evolution of microRNA genes in Drosophila species. Genome Biol Evol 2:180-189.

Ohshima K, Hattori M, Yada T, Gojobori T, Sakaki Y, Okada N. 2003. Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. Genome Biol 4:R74.

Ohta T. 1993. An examination of the generation-time effect on molecular evolution. Proc Natl Acad Sci U S A 90:10676-10680.

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

Ohta T, Kimura M. 1971. On the constancy of the evolutionary rate of cistrons. J Mol Evol 1:18-25.

Ophir R, Itoh T, Graur D, Gojobori T. 1999. A simple method for estimating the intensity of purifying selection in protein-coding genes. Molecular biology and evolution 16:49-53.

Pacifici M, Santini L, Marco DM, Baisero D, Francucci L, Marasini GG, Visconti M, Rondini M. 2013. Generation length in mammals. Nature Conservation:89-94.

Pagani F, Baralle FE. 2004. Genomic variants in exons and introns: identifying the splicing spoilers. Nature reviews. Genetics 5:389-396.

Parmley JL, Chamary JV, Hurst LD. 2006. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. Mol Biol Evol 23:301-309.

Pei B, Sisu C, Frankish A, Howald C, Habegger L, Mu XJ, Harte R, Balasubramanian S, Tanzer A, Diekhans M, et al. 2012. The GENCODE pseudogene resource. Genome Biol 13:R51.

Piganeau G, Eyre-Walker A. 2009. Evidence for variation in the effective population size of animal mitochondrial DNA. PLoS One 4:e4396.

Popadin KY, Nikolaev SI, Junier T, Baranova M, Antonarakis SE. 2013. Purifying selection in mammalian mitochondrial protein-coding genes is highly effective and congruent with evolution of nuclear genes. Mol Biol Evol 30:347-355.

Pozzoli U, Menozzi G, Comi GP, Cagliani R, Bresolin N, Sironi M. 2007. Intron size in mammals: complexity comes to terms with economy. Trends Genet 23:20-24.

Prasad AB, Allard MW, Green ED. 2008. Confirming the phylogeny of mammals by use of large comparative sequence data sets. Mol Biol Evol 25:1795-1808.

Proulx SR, Phillips PC. 2005. The opportunity for canalization and the evolution of genetic networks. Am Nat 165:147-162.

R Development Core Team. 2009. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Reddy BV, Datta S, Tiwari S. 1998. Use of propensities of amino acids to the local structural environments to understand effect of substitution mutations on protein stability. Protein Eng 11:1137-1145.

Rice SH. 1998. The evolution of canalization and the breaking of von baer's laws: Modeling the evolution of development with epistasis. Evolution 52:647-656.

Ritter AD, Shen Y, Fuxman Bass J, Jeyaraj S, Deplancke B, Mukhopadhyay A, Xu J, Driscoll M, Tissenbaum HA, Walhout AJ. 2013. Complex expression dynamics and robustness in C. elegans insulin networks. Genome Res 23:954-965.

Robinson DM, Jones DT, Kishino H, Goldman N, Thorne JL. 2003. Protein evolution with dependence among codons due to tertiary structure. Mol Biol Evol 20:1692-1704.

Rosenberg MS, Subramanian S, Kumar S. 2003. Patterns of transitional mutation biases within and among mammalian genomes. Molecular biology and evolution 20:988-993.

Rosenbloom KR, Dreszer TR, Pheasant M, Barber GP, Meyer LR, Pohl A, Raney BJ, Wang T, Hinrichs AS, Zweig AS, et al. 2010. ENCODE whole-genome data in the UCSC Genome Browser. Nucleic Acids Res 38:D620-625.

Rouchka C, Cha E. 2009. Current Trends in Pseudogene Detection and Characterization. Current Bioinformatics 4:112-119.

Sanjuan R, Cuevas JM, Furio V, Holmes EC, Moya A. 2007. Selection for robustness in mutagenized RNA viruses. PLoS Genet 3:e93.

Schneider A, Dessimoz C, Gonnet GH. 2007. OMA Browser--exploring orthologous relations across 352 complete genomes. Bioinformatics 23:2180-2182.

Schneider A, Souvorov A, Sabath N, Landan G, Gonnet GH, Graur D. 2009. Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. Genome Biol Evol 1:114-118.

Shabalina SA, Ogurtsov AY, Spiridonov NA. 2006. A periodic pattern of mRNA secondary structure created by the genetic code. Nucleic Acids Res 34:2428-2437.

Shen LX, Basilion JP, Stanton VP, Jr. 1999. Single-nucleotide polymorphisms can cause different structural folds of mRNA. Proc Natl Acad Sci U S A 96:7871-7876.

Shu W, Bo X, Ni M, Zheng Z, Wang S. 2007. In silico genetic robustness analysis of microRNA secondary structures: potential evidence of congruent evolution in microRNA. BMC Evol Biol 7:223.

Shu W, Ni M, Bo X, Zheng Z, Wang S. 2008. In silico genetic robustness analysis of secondary structural elements in the miRNA gene. J Mol Evol 67:560-569.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15:1034-1050.

Skelly DA, Ronald J, Connelly CF, Akey JM. 2009. Population genomics of intron splicing in 38 Saccharomyces cerevisiae genome sequences. Genome Biol Evol 1:466-478.

Smit A, Hubley, R & Green, P. 1996-2012. RepeatMasker Open-3.0.

Smith NG, Hurst LD. 1998. Sensitivity of patterns of molecular evolution to alterations in methodology: a critique of Hughes and Yeager. J Mol Evol 47:493-500.

Stelling J, Sauer U, Szallasi Z, Doyle FJ, 3rd, Doyle J. 2004. Robustness of cellular functions. Cell 118:675-685.

Stergachis AB, Haugen E, Shafer A, Fu W, Vernot B, Reynolds A, Raubitschek A, Ziegler S, LeProust EM, Akey JM, et al. 2013. Exonic transcription factor binding directs codon choice and affects protein evolution. Science 342:1367-1372.

Stoletzki N, Eyre-Walker A. 2007. Synonymous codon usage in Escherichia coli: selection for translational accuracy. Mol Biol Evol 24:374-381.

Svensson O, Arvestad L, Lagergren J. 2006. Genome-wide survey for biologically functional pseudogenes. PLoS Comput Biol 2:e46.

Szollosi GJ, Derenyi I. 2009. Congruent evolution of genetic and environmental robustness in micro-RNA. Mol Biol Evol 26:867-874.

Szollosi GJ, Derenyi I. 2008. The effect of recombination on the neutral evolution of genetic robustness. Math Biosci 214:58-62.

Tamura K, Subramanian S, Kumar S. 2004. Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks. Mol Biol Evol 21:36-44.

Terai G, Yoshizawa A, Okida H, Asai K, Mituyama T. 2010. Discovery of short pseudogenes derived from messenger RNAs. Nucleic acids research 38:1163-1171.

The Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69-87.

Torrents D, Suyama M, Zdobnov E, Bork P. 2003. A genome-wide survey of human pseudogenes. Genome Res 13:2559-2567.

Tsagkogeorga G, Parker J, Stupka E, Cotton JA, Rossiter SJ. 2013. Phylogenomic analyses elucidate the evolutionary relationships of bats. Curr Biol 23:2262-2267.

van Baren MJ, Brent MR. 2006. Iterative gene prediction and pseudogene removal improves genome annotation. Genome Res 16:678-685.

van Nimwegen E, Crutchfield JP, Huynen M. 1999. Neutral evolution of mutational robustness. Proc Natl Acad Sci U S A 96:9716-9720.

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 103:3220-3225.

von Dassow G, Meir E, Munro EM, Odell GM. 2000. The segment polarity network is a robust developmental module. Nature 406:188-192.

Wagner A. 2005. Robustness and evolvability in living systems. Princeton, N.J.: Princeton University Press.

Wagner A, Stadler PF. 1999. Viral RNA and evolved mutational robustness. J Exp Zool 285:119-127.

Wagner G, Booth G, Bagheri-Chaichian H. 1997. A population genetic theory of canalization. Evolution 51:329-347.

Wilke CO, Drummond DA. 2006. Population genetics of translational robustness. Genetics 173:473-481.

Wilke CO, Wang JL, Ofria C, Lenski RE, Adami C. 2001. Evolution of digital organisms at high mutation rates leads to survival of the flattest. Nature 412:331-333.

Wright S. 1931. Evolution in Mendelian Populations. Genetics 16:97-159.

Wright SI, Yau CB, Looseley M, Meyers BC. 2004. Effects of gene expression on molecular evolution in Arabidopsis thaliana and Arabidopsis lyrata. Mol Biol Evol 21:1719-1726.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Computer applications in the biosciences : CABIOS 13:555-556.

Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol 17:32-43.

Yang Z, Nielsen R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Biol 46:409-418.

Zeng Y, Cullen BR. 2005. Efficient processing of primary microRNA hairpins by Drosha requires flanking nonstructured RNA sequences. J Biol Chem 280:27595-27603.

Zeng Y, Cullen BR. 2003. Sequence requirements for micro RNA processing and function in human cells. RNA 9:112-123.

Zeng Y, Cullen BR. 2004. Structural requirements for pre-microRNA binding and nuclear export by Exportin 5. Nucleic Acids Res 32:4776-4785.

Zeng Y, Yi R, Cullen BR. 2005. Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha. EMBO J 24:138-148.

Zhang Z, Gerstein M. 2004. Large-scale analysis of pseudogenes in the human genome. Curr Opin Genet Dev 14:328-335.

Zhang Z, Harrison P, Gerstein M. 2002. Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. Genome Res 12:1466-1482.

Zhang Z, Harrison PM, Liu Y, Gerstein M. 2003. Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. Genome Res 13:2541-2558.

Zhou T, Gu W, Wilke CO. 2010. Detecting positive and purifying selection at synonymous sites in yeast and worm. Mol Biol Evol 27:1912-1922.

Zimin A, Cornish A, Maudhoo M, Zhang X, Pandey S, Wipfler K, Marçais G, Roberts M, Ferguson B, Grandnigo SJ, et al. 2013. Rhesus macaque genome 2.0. Not Puplished.

Zuker M, Stiegler P. 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res 9:133-148.