Structural and population genetic determinants of RNA secondary structure evolution

Robert Piskol

München, 2011

Structural and population genetic determinants of RNA secondary structure evolution

Dissertation

zur
Erlangung des Doktorgrades der Naturwissenschaften
der Fakultät für Biologie an der Ludwig-Maximilians-Universität München

vorgelegt von
Robert Piskol
aus Brieg

München, 2011

Erstgutachter: Prof. Dr. Wolfgang Stephan
Zweitgutachter: Prof. Dr. John Parsch

Tag der Abgabe: 24.03.2011
Tag der mündlichen Prüfung: 26.04.2011

Ehrenwörtliche Versicherung und Erklärung

Diese Promotion wurde im Sinne des §12 der Promotionsordnung von Prof. Dr. Wolfgang Stephan betreut.

Hiermit erkläre ich, dass die vorliegende Dissertation von mir selbständig und ohne unerlaubte Hilfe angefertigt wurde. Zudem wurden keine anderen als die angegebenen Quellen verwendet.

Ich versichere, dass die Dissertation keiner anderen Prüfungskommission vorgelegt wurde und ich mich nicht anderweitig einer Doktorprüfung ohne Erfolg unterzogen habe.

München, 24.03.2011

Robert Piskol

Erklärung über Beitrag als Author

In dieser Dissertation präsentiere ich die Ergebnisse meiner Doktorarbeit, die von Januar 2008 bis März 2011 durchgeführt wurde. Der Kern dieser Arbeit besteht aus drei Kapiteln. Zwei dieser Kapitel sind als Publikationen in einer international anerkannten Fachzeitschrift erschienen, während ein Kapitel momentan unter Begutachtung steht. Die Studien in diesen Kapiteln wurden von Wolfgang Stephan und mir konzipiert. Ich führte die Studien durch und analysierte die Daten. Die Manuskripte wurden von mir und Wolfgang Stephan verfasst.

Robert Piskol

Wolfgang Stephan

Zusammenfassung

Seit der Entdeckung der RNA, werden unaufhörlich neue Funktionen gefunden, die dieses Molekül vollführt. Sie gehen weit über dessen ursprüngliche Rolle als Übermittler von Information in der Biosynthese von Proteinen hinaus, durch die es seine Bekanntheit erlangt hat. Solche nichtkodierenden RNAs (ncRNAs) sind an fundamentalen zellulären Prozessen, wie zum Beispiel der Regulation von Genexpression und Erhaltung der Genomstabilität, beteiligt. In vielen Fällen ist die Biogenese oder Funktion der RNA nur möglich, wenn das Molekül eine charakteristische zwei- und dreidimensionale Faltung annimmt, die durch Bildung von intramolekularen Basenpaaren entsteht. Die Trennung dieser paarenden Regionen durch Mutationen in der Sequenz kann zu Veränderungen in der Konformation des Moleküls führen, die eine potentielle Beeinträchtigung der korrekten Funktion nach sich zieht. Sogenannte 'kompensatorische' Mutationen haben jedoch die Fähigkeit das Molekül in seine ursprüngliche Konformation zurück zu führen. Unter dem Einfluss verschiedener evolutionärer Kräfte häufen sich diese Doppelmutationen (Kovariationen) in paarenden Regionen des RNA Moleküls (sog. Helices) an. Die Wahrscheinlichkeit von Kovariationen, und damit die Rate der Evolution hängt von verschiedenen Eigenschaften der Helix ab.

Durch Verwendung eines Ansatzes, der auf logistischer Regression beruht, war es uns möglich, die evolutionären Dynamiken in RNA Molekülen zu studieren (Piskol and Stephan, 2008). Diese Methode wurde auf einen Datensatz von vorhergesagten RNA Sekundärstrukturen in Vertebratenintrons angewandt. Unser Ziel war die Erforschung von strukturellen und populationsgenetischen Faktoren, die die Rate von kompensatorischen Mutationen in RNA Molekülen beeinflussen. Wie von Kimura's (1985) Modell von kompensatorischer Evolution vorhergesagt, sind unsere Ergebnisse mit der Hypothese vereinbar, dass die physikalische Distanz zwischen paarenden Nukleotiden einen negativen Einfluss auf das Auftreten von Kovariationen hat. Weiterhin konnte festgestellt werden, dass längere paarende Regionen eine größere Anzahl von 'wobbles' (GU Basenpaare) und 'mismatches' (nicht gepaarte Nukleotide) tolerieren können und letztendlich auch mehr Kovariationen enthalten. Darüber hinaus enthüllte die positionsweise Analyse aller Nukleotide in paarenden Regionen, dass Kovariationen bevorzugt in den äußeren Bereichen einer Helix auftreten, während 'wobbles' und 'mismatches' häufiger an weiter innenliegenden Positionen anzutreffen sind. Bei diesem Prozess scheint der Gehalt von Guanin und Cytosin eine erhebliche Rolle zu spielen.

Die oben verwendeten Daten wurden durch RNA Moleküle erweitert, die in den kom-

pletten nuklearen Genomen von Drosophiliden (*Drosophila melanogaster/D. simulans*) und Hominiden (Mensch/Schimpanse) anzufinden sind. Wir bestimmten genomweite selektive Zwänge für diese Moleküle (Piskol and Stephan, 2011). Im Vergleich zu neutral evolvierenden Regionen der betrachteten Genome, fanden wir stark reduzierte Substitutionsraten an gepaarten und ungepaarten Positionen in gefalteten Molekülen. Wir berechneten, dass mehr als 90% aller neuen Mutationen in nichtkodierenden RNAs durch purifizierende Selektion aus der Sequenz entfernt wurden. Diese Werte übersteigen Schätzungen an genomischen Positionen, die zu einem Aminosäureaustausch führen können und stellen die Bedeutsamkeit vieler gefalteter genomischer Regionen und ihrer Funktion heraus (korrektes Spleißen, Effizienz des Spleißens, Lokalisierung von Proteinen, RNA Modifikation). Wir konnten keine signifikanten Unterschiede in selektiven Einschränkungen finden, die durch die genomische Postition der Moleküle bedingt wären (kodierend/nichtkodirerend, genisch/intergenisch, UTR/non-UTR). Deswegen, scheint die eingeschränkte Evolution von ncRNAs vor allem durch das grundlegende Bedürfnis zur Erhaltung der Paarung von Nukleotiden bestimmt zu sein, und nur im geringen Umfang durch die genomische Position des Moleküls beeinflusst zu werden. Aus dem Vergleich von Selektionskoeffizienten zwischen Drosophiliden und Hominiden wurde zudem ersichtlich, dass auch die effektive Populationsgröße einen Einfluss auf die Evolution von RNA Molekülen hat. Ihre Wirkung resultiert in signifikant höheren Einschränkungen in Drosophiliden und beeinflusst Evolution stärker an ungepaarten Positionen.

Motiviert durch Anzeichen für die Rolle der effektiven Populationsgröße in der Evolution von RNA Molekülen erforschten wir dieses Thema im Detail. Die effektive Populationsgröße einer Spezies (N_e) ist eine fundamentale Einheit in der Populationsgenetik. Ihr Einfluss auf die Wirksamkeit von Selektion ist Bestandteil vieler theoretischer und empirischer Studien. Jedoch wurde der Effekt von N_e meist im Zusammenhang mit der Evolution von unabhängig evolvierenden genomischen Postionen betrachtet, während ihr Einfluss auf epistatische Interaktionen (Interaktionen zwischen zwei oder mehreren genomischen Positionen) weitgehend unklar ist. Unsere vorherige Arbeit belegte die Rolle von N_e in der Evolution von RNA Molekülen (welche zu großen Teilen aus koevolvierenden Regionen bestehen). Um unser momentanes Wissen über den Einfluss von N_e auf unabhängig evolvierende und koevolvierende genomische Positionen zu erweitern, fokussierten wir unsere Arbeit auf transfer RNAs (tRNAs) – eine Klasse von RNA Molekülen mit wohl bekannter Struktur und Funktion. Wir verglichen die evolutionären Raten an gepaarten und ungepaarten Positionen in orthologen tRNAs verschiedener Paare von Vertebraten- und Drosophilaspezies. Hierfür wählten wir Gruppen von Spezies die sich in ihrer langfristigen effektiven Populationsgröße unterscheiden und verglichen ihre Level von selektiven Zwängen. Die verwendeten Speziespaare waren Mensch/Makak, Makak/Krallenaffe, Hund/Katze, Huhn/Zebrafink, Maus/Ratte, *D. melanogaster*/*D. yakuba*, *D. melanogaster*/*D. simulans*. In der Tat können Differenzen in selektiven Einschränkungen aufgrund von Unterschieden in N_e beobachtet werden. Dieser Effekt ist stärker an ungepaarten (unabhängig evolvierenden) Positionen als an gepaarten (koevolvierenden) Postionen. Weiterhin konnten wir für alle Speziespaare orthologe tRNAs anhand ihrer Ähnlichkeit zu tRNAs in allen Reichen des Lebens in ein Kernset (mit hoher Ähnlichkeit) und ein peripheres Set (mit geringer Ähnlichkeit) aufteilen. Wir stellten dabei fest, dass tRNAs im Kernset stärkeren evolutionären Ein-

schränkungen ausgesetzt sind und nur unter einem schwachen Einfluss von N_e stehen, während tRNAs im peripheren Set einer starken Wirkung von N_e unterliegen. Wir überprüften auch die Wirkung von N_e, die durch ungleiche effektive Populationsgrößen zwischen Autosomen und X Chromosomen in Drosophiliden hervorgerufen werden können und stellten auch hier fest, dass evolutionäre Einschränkungen für tRNAs auf dem X Chromosom gelockert sind, was durch die geringere effektive Populationsgröße für das X Chromosom erklärt werden kann.

Summary

Since their discovery, RNA molecules have been shown to carry functions that extend far beyond their initially ascribed role as intermediates in protein biosynthesis. These noncoding RNAs (ncRNAs) are involved in fundamental cellular processes including the regulation of gene expression and maintenance of genome stability. In most cases the biogenesis or function of the RNA molecule is only possible if the molecule folds into a characteristic two- and three-dimensional shape via formation of intra-molecular base pairs. The disruption of these paired regions through mutations in the primary sequence can result in conformational changes of the molecule that impair its ability to function correctly. However, compensatory mutations can restore the original conformation of the molecule. Under the influence of various evolutionary forces, such as mutation and selection, a paired region (helix) will accumulate these nucleotide double-substitutions (covariations). The chance of a substitution and thus the rate of evolution depends on different properties of the helix.

We developed a logistic regression approach to analyze the evolutionary dynamics of RNA secondary structures (Piskol and Stephan, 2008). This method was applied to a set of computationally predicted RNA secondary structures in vertebrate introns. Our aim was to discover structural and population genetic determinants of the compensatory mutation rate in RNA molecules. As predicted by Kimura's (1985) model of compensatory evolution, our results are in agreement with the hypothesis that the physical distance between pairing nucleotides has a negative influence on the occurrence of covariations. Furthermore, we found that longer pairing regions have the ability to tolerate more wobbles (GU base pairs) and mismatches, and ultimately also contain more covariations. The position-wise analysis of all nucleotides in paired regions revealed that covariations occur preferentially at the helix ends, whereas wobbles and mismatches are more frequent in the middle of a helix. This pattern is largely determined by the GC content.

We extended the study described above from structured regions in introns of vertebrate genes to folded RNA molecules that are scattered across the whole nuclear genomes of drosophilids (*Drosophila melanogaster*/*D. simulans*) and hominids (human/chimp). For these molecules we estimated genome wide selective constraints (Piskol and Stephan, 2011). In comparison to neutrally evolving regions of the same genomes we observed substantially reduced rates of substitutions at paired and unpaired sites of folded molecules.

We found that more than 90% of novel mutations in ncRNAs are removed from the sequence by purifying selection. These values exceed estimates that were previously obtained for amino-acid changing positions of protein coding genes. It points to the overall importance of many folded genomic regions, which carry quite diverse functions (correct splicing, splicing efficiency, protein localization, RNA editing). We did not find significant differences in constraints between folded molecules based on their genomic location (coding/noncoding, genic/intergenic, UTR/non-UTR). Therefore, the restricted evolution of ncRNAs seems to be mostly driven by the basic need of the molecule to remain in its original conformation through continuous maintenance of pairings between nucleotides and only to a smaller extent by the location of the molecule in the genome. In addition, a comparison of selective coefficients between drosophilids and hominids enabled us to find evidence for the impact of the effective population size on RNA evolution, which resulted in significantly higher constraints in drosophilids than hominids and led to larger differences in selective constraints at unpaired than at paired positions.

Motivated by the evidence for a potential role of the effective population size in the evolution of ncRNA molecules we explored this topic in greater detail. The effective population size of a species (N_e) is a fundamental quantity in population genetics. Its impact on the efficacy of selection has been the focus of many theoretical and empirical studies over the recent years. Yet, the effect of N_e was mostly investigated in connection with the evolution of independently evolving sites in a genome, while its impact on the evolution of epistatic interactions is not well understood. Our previous work (see previous paragraph) showed evidence for the role of N_e in the evolution of ncRNA molecules (which consist to a large extent of coevolving regions). To increase our knowledge of the impact of N_e on the evolution at independently evolving and coevolving sites, we focused on transfer RNAs (tRNAs) – a class of RNA molecules with well studied structure and function. We compared the rates of evolution at paired and unpaired positions in orthologous tRNAs of various vertebrate and Drosophila species. Therefore, we chose groups of species that differ in their long-term effective population sizes and compared the level of selective constraint between them. These pairs included human/macaque, macaque/marmoset, dog/cat, chicken/zebra finch, mouse/rat, *D. melanogaster*/*D. yakuba*, and *D. melanogaster*/*D. simulans*. Indeed, we were able to detect differences in selective constraints between species pairs of different N_e. These differences can be explained well by theoretical predictions for the evolution of independently evolving and coevolving sites. Specifically, we found that constraints in orthologous tRNAs of a species pair increase with increasing long-term N_e. Thereby, the effect of N_e is stronger at unpaired (independently evolving) sites than paired (coevolving) sites. Furthermore, for all species pairs we identified sets of orthologous tRNAs with high structural similarity to tRNAs from all major kingdoms of life ('core' sets), and tRNAs with lower similarity ('peripheral' sets). We found the core sets to be under strong overall constraints and only subject to a negligible effect of N_e. In the peripheral set, however, we discovered a strong influence of N_e on constraints. We also investigated whether the difference in N_e between autosomes and X chromosome, due to the presence of the X chromosome in one copy in males, has an effect on differences in evolutionary rates. We were able to show that constraints are more relaxed in X-linked tRNAs.

List of Publications

The work during this doctoral thesis has resulted in two publications and one accepted manuscript that is currently under minor revision. They constitute chapters 2, 3, and 4 of this thesis and are supplemented by appendices A, B, and C:

- Piskol R, Stephan W. 2008. Analyzing the evolution of RNA secondary structures in vertebrate introns using Kimura's model of compensatory fitness interactions. *Mol. Biol. Evol.* 25(11):2483–2492.

- Piskol R, Stephan W. 2011. Selective constraints in conserved folded RNAs of drosophilid and hominid genomes. *Mol. Biol. Evol.* 28(4):1519–1529

- Piskol R, Stephan W. 2011. The role of the effective population size in compensatory evolution. *Genome Biol. Evol.* (accepted with minor revision)

Contents

List of Figures

List of Tables

Chapter 1

General Introduction

1.1 The Biological Roles of RNAs

The sequencing of the whole human genome has revealed that protein-coding regions account for only 1.5% of the complete genomic sequence (Lander et al., 2001). Nevertheless, a substantial fraction of the genome is transcribed and results in large numbers of noncoding ribonucleic acids (ncRNAs) (Mattick, 2009). These ncRNAs are key players in a multitude of biological processes and have extended our notion of the RNA as the sole carrier of information between DNA and proteins (the central dogma of molecular biology, Crick (1970)) and as parts of the protein production machinery in the form of transfer RNA (tRNA) and ribosomal RNA (rRNA). The pool of ncRNAs in an eukaryotic cell comprises a great variety of RNA molecules that are as diverse in lengths and shapes as they are in their functions. Well known members of long ncRNAs (>200 nucleotides) in that pool are *Air*, *HOTAIR*, and *Xist*. Each of them is several kilobases long and involved in epigenetic silencing as well as X chromosome inactivation (Duret et al., 2006; Nagano et al., 2008; Rinn et al., 2007), respectively. Many of these long ncRNAs are only poorly conserved between species (Pang et al., 2006). However, there also exist large numbers of small RNAs with high conservation. These include, among others, small nucleolar RNAs (snoRNAs), piwi-interacting RNAs (piRNAs), small interfering RNAs (siRNAs), and micro RNAs (miRNAs) (Amaral et al., 2008). They are involved in catalytic modification of rRNAs, chromatin state regulation, and the regulation of cell proliferation and development (Galasso et al., 2010). The expression of siRNAs and miRNAs usually depends on

the tissue and developmental stage (Gutierrez-Aguilar et al., 2010) and is responsible for specific posttranscriptional regulation of gene expression (Hobert, 2008) and gene silencing (Meister and Tuschl, 2004). Even though different in sequence and function, many of these ncRNA molecules have in common that their biogenesis, functional efficiency, or both depend on the structural conformation of the molecule. Similarly to DNA, RNA also consists of complementary nucleotides that have the ability to establish nucleotide pairs through hydrogen bonds. However, in contrast to the double stranded nature of DNA in which complementarity occurs between two molecules of the same type, the single-stranded state of an RNA molecule facilitates nucleotide pairings within one molecule that lead to distinct two- and three-dimensional conformations. These structures can occur on the whole length of the molecule (tRNAs, snoRNAs, precursors of miRNAs and siRNAs), or only locally. Examples of locally formed structured regions include 1) paired regions in introns of protein-coding genes that are important for the correct inclusion or exclusion of exons (Howe and Ares, 1997), 2) the regulation of splicing efficiency (Chen and Stephan, 2003), or 3) correct localization of mRNA transcripts through structures in 3' untranslated regions (UTRs) (Bullock et al., 2003; MacDonald, 1990; Irion and Johnston, 2007).

In many of these cases nucleotide variation in the structured regions has been demonstrated to be associated with altered or aberrant phenotypes, which underlines the importance of the folded structure of these molecules. Often the change of an organism's phenotype also results in a change of its fitness. Therefore, mutations in regions that alter the secondary structure may have positive or negative consequences on the reproductive success of an individual. Specifically, if we assume that organisms (and their biological processes) are overall already well adapted to their current environment, the proportion of mutations that confer a reduction in fitness will be much greater than the chance that a mutation is of a beneficial nature (Silander et al., 2007). Therefore, one can assume that RNA molecules will have to cope mostly with mutations that confer a reduction in fitness.

While the structure and function of many ncRNAs and structured regions in protein-coding transcripts has been elucidated, the pool of unknown ncRNAs is by far not yet exhausted, as can be seen by the continuous discoveries of new ncRNA classes (e.g., tran-

scription initiation RNAs (tiRNAs) (Taft et al., 2009), splice-site RNAs (spliRNAs) (Taft et al., 2010*a*), and telomere-specific small RNAs (tel-sRNAs) (Cao et al., 2009)). These findings have been facilitated by the recent availability of whole-genome data for many model species (Mural et al., 2002; Chimpanzee Sequencing and Analysis Consortium, 2005; Drosophila 12 Genomes Consortium, 2007) and the advent of high-throughput sequencing techniques that allow the efficient characterization of whole transcriptomes. The combination of comparative genomics and computational methods allows the effective screening for conserved ncRNAs, as it was demonstrated by recent studies in Drosophila (Rose et al., 2007; Stark et al., 2007; Bradley et al., 2009) and vertebrates (Pedersen et al., 2006; Washietl et al., 2007). The exploration of the available information has also been greatly simplified by integral online resources including mirBase (Griffiths-Jones et al., 2008) and Rfam (Gardner et al., 2009), which store all experimentally verified miRNAs and organize known ncRNAs into RNA families, respectively.

1.2 Comparative Genomics and Population Genetics of RNA Molecules

The wealth of genomic data available today allows for large scale comparison of orthologous RNAs between species. Given that the phylogenetic relationship between the investigated organisms is known, we are able to trace changes at single nucleotide positions in the genomic sequence back to the ancestor of these species, which allows us to make conclusions about the evolutionary history of the region of interest. It is important to realize that the nucleotide variation, which is observed between different species is the result of processes that occurred at a population level. Thereby, mutations are occurring with a certain rate at a genomic locus in a population. The fate of these mutations is determined by their selective advantage (or disadvantage) as well as their random sampling from one generation to the next (genetic drift). Three types of selection are possible. Selection can be positive, and thus lead to a faster fixation of alleles in a population. Also purifying or to a lesser extent balancing selection are possible, which speed up the loss of mutant alleles, or stabilize the frequency of an allele in a population, respectively. Therefore, mutation (the source of nucleotide variation), selection and genetic drift influence the

mutant allele frequency in a population and may eventually lead to its rise in frequency such that all individuals of a species carry that nucleotide variant. This so-called "fixation event" results in a nucleotide difference that can be observed between the sequences of two species. Therefore, population genetics can supply us with the necessary theory to study and interpret nucleotide divergence patterns and also provides hypotheses that can be tested using inter-species comparisons.

One of the population genetic key concepts, whose implications extend into comparative genomics, is the time until fixation of a mutant allele in the population. Pioneered by the work of Kimura for selectively neutral mutations (Kimura, 1983), this idea has also been applied to mutations that are selected against (Kimura, 1962, 1980). This theory builds on the irreversibility of mutations and assumes independently evolving nucleotides. A more realistic treatment that takes back mutations into account is possible through simulations of the Wright-Fisher process using a pseudo sampling variable (PSV) (Kimura and Takahata, 1983). This procedure simulates the diffusion process of mutations in a population, which experience drift and selection. While many regions in the genome evolve independently (i.e., the change in frequency at a given site is independent of the state of other sites) and fixation times at these sites can be described by the above methods, these models do not apply to RNA molecules in general. Due to its composition of unpaired nucleotides as well as Watson-Crick (WC) nucleotide pairs[1] (Figure 1.1) parts of the RNA sequence may evolve independently, while other regions are subject to coevolutionary dynamics. Therefore, mutations in paired regions behave differently than mutant alleles at independently evolving positions. Considering a pair of nucleotides that are involved in a WC pair, a single mutation will inevitably disrupt the pairing and confer a reduction in fitness, while a second so called "compensatory" mutation at the opposing position has the ability to restore the original pairing and thus can also restore the fitness of that pair (Figure 1.1). In this process the sequence of the RNA will change at two positions while the structure of the initial and final states remains the same. Fixation times for such double mutations have been obtained analytically in the case of a strong reduction in fitness and irreversible (Kimura, 1985; Stephan, 1996) or reversible (Higgs, 1998) mutations, and in the case when selection against intermediate configurations is weak (Innan and Stephan,

[1]Nucleotide pairs between guanine and cytosine (GC) or adenine and uracil (AU)

2001). The fixation of mutant nucleotides within the pair can occur either simultaneously (also called stochastic tunneling (Iwasa et al., 2004)) or sequentially, depending on whether selection against the intermediate nucleotide configuration is strong or weak, respectively.

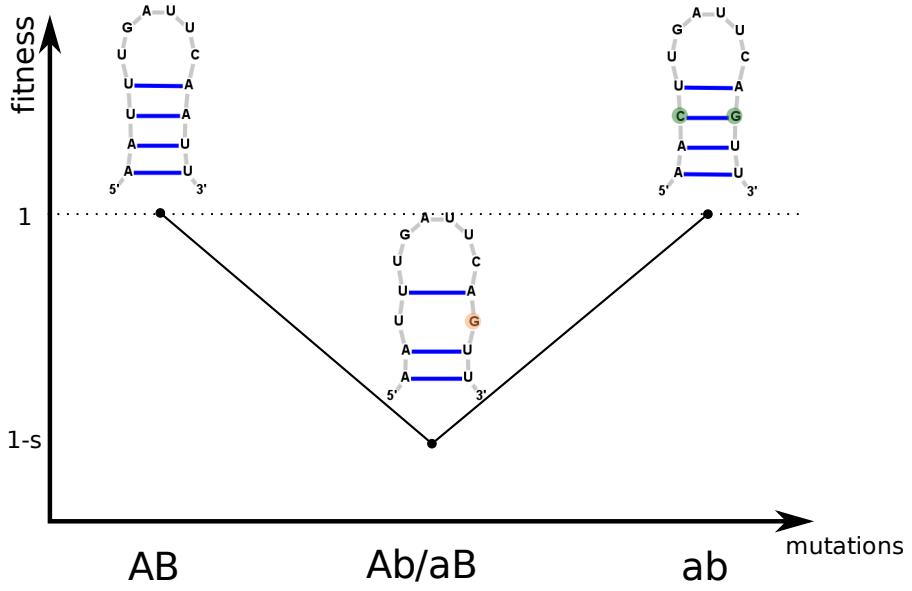

FIGURE 1.1. **Kimura's (1985) two-locus model of compensatory evolution.** Individual mutations lead to intermediate conformations (Ab/aB) that confer a reduction of fitness (s), while initial (AB) and terminal (ab) states of the compensation process are neutral. Here, a hypothetic RNA molecule is depicted, where the backbone is colored in gray, whereas blue bars symbolize canonical Watson-Crick base pairs.

An important factor in the consideration of fixation events that involve two loci is the linkage between them. While mutations in close proximity to each other will be rarely separated after their occurrence, mutations separated by a greater distance will be subject to larger rates of recombination in between, which might lead to the separation of beneficial combinations (Stephan and Kirby, 1993; Kirby et al., 1995). Therefore, the fixation probability of a double mutation decreases with growing distance between paired nucleotides if selection against intermediates (disrupted pairs) is strong ("distance effect"). However, it has also been shown that in other regimes of selection and recombination, crossing over may create new neutral, or even beneficial combinations (Lynch, 2010; Weissman et al., 2010).

1.3 RNA Structure Evolution

The three-dimensional appearance of the RNA molecule is organized on two levels into secondary and tertiary structure. The secondary structure of the molecule arises through contacts between parts of the RNA nucleotide chain that are complementary to each other and allow the formation of WC pairs between nucleotides. The resulting local contacts between antiparallel parts of the backbone are called "stems" or "helices", as they are similar to the double stranded nature of a DNA double helix. It is believed that secondary structures are the first to form during the folding of the molecule (Zarrinkar and Williamson, 1994; Tinoco and Bustamante, 1999). The three-dimensional form of the molecule is achieved through additional nucleotide pairs that can involve not only the Watson-Crick edge of the nucleotide but also the Hoogsteen- or sugar-edge and can be either in cis or trans orientation of the glycosidic bonds (see Figure 1.2). While usually no coevolution of these non-WC pairs is observed (Dutheil et al., 2010), canonical WC pairs are subject to coevolutionary dynamics (Chen and Stephan, 2003), and can therefore be used as indicators for pairings between regions if a compensatory mutation is observed. Compensatory mutations have been proposed to be the prevalent mode of RNA evolution due to the observation of increased levels of linkage disequilibrium (nonrandom association between loci) in paired regions of the Drosophila *alcohol dehydrogenase* (*Adh*) gene (Chen et al., 1999) and a clustering of compensatory mutations on the same terminal branches of the phylogenetic tree for mitochondrial tRNAs (Meer et al., 2010).

In this respect, it is of interest to understand at which rate coevolutionary events in structured RNA molecules occur and whether this rate is regulated by certain factors. We may imagine that the rate will depend on population genetic forces but also structural parameters of the molecule. Population genetic parameters include the effective population size (which will be discussed in the next section in greater detail) and the recombination rate between pairing loci, which can be quantified by the physical distance (in nucleotides) between the two positions and was found to be negatively correlated with the rate of compensatory evolution. Structural parameters include the length of pairing regions and the position of a pair in the pairing region. Parsch et al. (2000) found that the rate of compensatory evolution is positively correlated with the helix length in

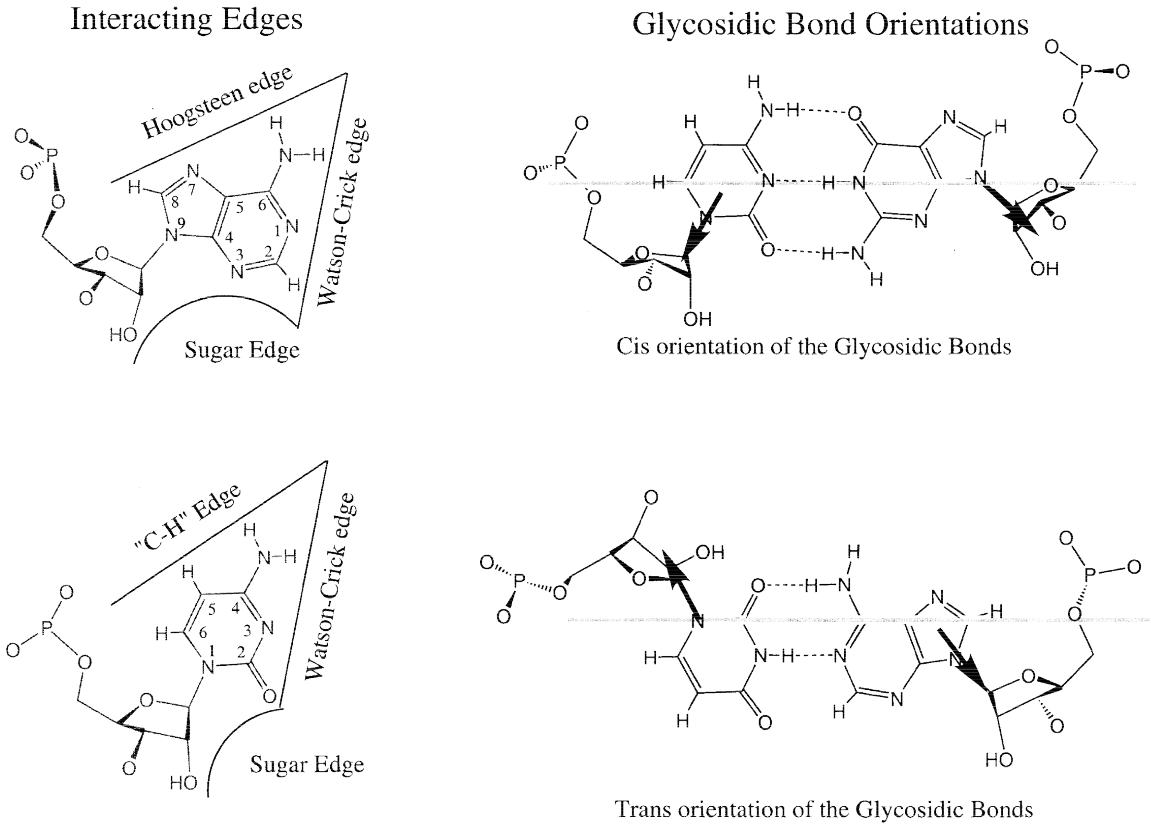

FIGURE 1.2. **Interacting edges for purines (top left) and pyrimidines (bottom left) as well as cis and trans orientations of glycosidic bonds.** A canonical Watson-Crick (WC) base pair involves hydrogen bonds between the WC edges of adenine and uracil or guanine and cytosine in cis-conformation, respectively (as shown for cytosine and guanine bases in the top right panel). The lower right panel displays a WC/WC trans interaction between uracil and adenine bases. (Figure adapted from Leontis and Westhof (2001))

RNase P RNA and were able to verify the presence of the distance effect in the 3' UTR of the *bicoid* gene. Increased rates of compensatory evolution with growing helix length were also observed in ribosomal RNA (Dutheil et al., 2010) and are attributed to the increased tolerance of longer helices to accommodate more unfavorable nucleotide combinations. Based on thermodynamic and structural criteria for RNA molecules, Mimouni et al. (2009) found that the substitution rate (i.e., the per nucleotide rate at which substitutions between two species can be observed), was lowest at penultimate helix positions, intermediate at ultimate positions and highest for nucleotides that are separated by more than two positions from the end of the helix. This suggests that helices are an essential structural unit in RNA structure evolution and influence compensatory mutation rates in various manners. In addition, substitution rates vary across the human genome and

are affected by the GC content of the sequence (Smith et al., 2002; Előry et al., 2010). Furthermore, evolution through compensatory mutations was shown to proceed faster when the intermediate state has slightly reduced fitness. In the case of rRNAs a larger fraction of compensatory mutations proceeds through a wobble[2] intermediate (Rousset et al., 1991), which changes the conformation of the backbone only slightly and leads to a moderate reduction in fitness. Therefore, the process of compensation may also depend on the nucleotide composition of the RNA and may choose different trajectories through the fitness landscape.

1.4 Effective Population Size and its Role in Evolution

While the recombination rate may be an important population genetic parameter that influences rates of evolution in RNA molecules, also the long-term effective population size of a species (N_e) may play a crucial role. N_e affects the rate of evolutionary change through the random sampling of individuals from one generation to the next (genetic drift) and also influences the efficacy of selection (Charlesworth, 2009). Mutations are considered to be nearly neutral if their selective disadvantage (s) is small in comparison to the effective population size ($s \ll \frac{1}{2N_e}$). Their fixation in such cases is mostly dominated by genetic drift. Slightly deleterious mutations ($s \approx \frac{1}{2N_e}$) may also fix in the population though at a smaller rate, while fixation for strongly deleterious mutations ($s \gg \frac{1}{2N_e}$) is usually unlikely (Chamary et al., 2006). As a result, the classification of mutations depends on the effective population size of the species they occur in. Mutations with a disadvantage s may be nearly neutral in species of small N_e and characterized by short fixation times but deleterious in species of large N_e (Andolfatto et al., 2011; Előry et al., 2010). As a consequence purifying selection is more efficient in species with a large N_e and will lead to a smaller number of divergent nucleotide positions compared to species with small N_e. However, N_e does not influence the fixation process only through its contribution to the strength of selection ($N_e s$). Also the population mutation rate (θ) is dependent on N_e and is typically given as a scaled mutation rate ($\theta = 4N_e\mu$).

[2] a pairing of guanine and uracil that is stabilized by two hydrogen bonds

To date, empirical studies that reported evidence for the role of N_e have focused on independently evolving sites. While theoretical results suggest that N_e is of no relevance for the evolution of interacting sites if selection is strong (Stephan, 1996) and also theoretical treatments of this subject exist in the case of weak selection (Innan and Stephan, 2001), only an empirical study can show how the effect of N_e compares between independently evolving and coevolving sites in nature.

1.5 Selective Constraints in Eukaryotic Genomes

It has been shown that 3–5% of the mammalian genome are subject to purifying selection (Mouse Genome Sequencing Consortium, 2002; Lunter et al., 2006) and that constraints on sequence evolution differ between classes of genomic sites as well as between taxa (Koonin and Wolf, 2010). In hominids and murids, for instance, alternatively spliced genes show larger selective constraints than single-transcript genes. Furthermore, constraints constantly increase between intronic, 3' untranslated, 5' untranslated, synonymous, and nonsynonymous regions (Előry et al., 2010). Also increasing constraints can be observed with decreasing distance to the transcription start and end positions of genes in hominids and murids (Keightley and Gaffney, 2003).

Protein-coding regions contribute to ∼1.5% of the complete genetic material in mammals. Therefore, only a small part of the estimated portion of constrained genomic regions is attributable to them. The remaining constrained sites are believed to be located to a large extent in noncoding RNA genes. However, explicit characterizations of the strength of selection in RNA molecules have not been performed until now. Comparative genomics allows us to determine the strength of selection in ncRNAs located in different regions of the genome. The estimation of constraints can be performed by comparison of nucleotide variation in regions of interest to sequences that are assumed to have evolved neutrally. Usually the difficulty lies in the determination of a neutral standard of sequence evolution in the organism of interest. Due to the degeneracy of the genetic code many studies originally used synonymous sites as a standard for unconstrained evolution. However, it becomes more and more evident that synonymous sites are subject to selection as well. This is partially due to the selective advantage of codons that increase the speed or

accuracy of translation (codon bias) (Akashi, 1994; Stenoien et al., 2000; Stoletzki and Eyre-Walker, 2007) but also due to selection that maintains a specific secondary structure of the mRNA (Hoede et al., 2006; Stoletzki, 2008). As a result other genomic regions have been suggested to replace synonymous sites in their function as a standard for neutral evolution. In Drosophila it has been shown that positions 8–30 of short introns (<65nt) evolve in the most neutral fashion compared to other classes of sites (synonymous and nonsynonymous sites, long introns, etc.) (Parsch et al., 2010). In hominids and murids repetitive regions were suggested to serve as a neutral standard due to their higher rates of substitution compared to nonsynonymous and unique intronic/intergenic sites (Lunter et al., 2006; Előry et al., 2010). These sites can be used to calibrate the levels of observed divergence in regions of interest, which allows for the comparison of constraints between species.

1.6 Scope of this Dissertation

The aim of this thesis is to extend our understanding of evolutionary processes in RNA molecules. Therefore the following questions were covered within this thesis:

1. Which factors influence the rate of compensatory evolution?
2. How strong is purifying selection in ncRNAs?
3. Which role does the effective population size play in the evolution of RNA molecules?

Motivated by the fact that introns of protein-coding genes may harbor structured regions, which are responsible for the efficiency of splicing (Chen and Stephan, 2003), Chapter 2 describes the use of computationally predicted folded regions in vertebrate introns to determine influencing factors on the rate of compensatory evolution. The alignment of sequences from species that belong to a wide phylogenetic range allowed for the discovery of compensatory mutation events. Using a logistic regression framework the occurrence of these covariations was related to structural and population genetic parameters that characterized the molecule. In Chapter 3 the investigation of whole-genome annotations of folded RNA molecules in drosophilids (*Drosophila melanogaster*/*D. simulans*) and hominids (human/chimpanzee) gave insights into the selective pressures that are exerted on RNA molecules and demonstrates their functional importance. Thereby, substitution

patterns at paired and unpaired sites were interpreted from population and comparative genetics points of view. The comparison of drosophilids and hominids was of particular interest, since these two genera have substantially different long-term effective population sizes, which result in significantly different selective constraints. In CHAPTER 4, the inclusion of a larger number of vertebrate species and another Drosophila species (*D. yakuba*) allowed for a fine grained analysis of the role of the effective population size in the evolution of tRNAs with particular focus on independently evolving and coevolving sites. Thereby, not only the difference in effective population sizes between species was investigated, but also differences in N_e between X chromosome and autosomes within populations (here within *D. melanogaster* and *D. yakuba*) were examined.

Chapter 2

Analyzing the Evolution of RNA Secondary Structures in Vertebrate Introns Using Kimura's Model of Compensatory Fitness Interactions

Robert Piskol and Wolfgang Stephan (2008)
Mol. Biol. Evol., 25(11):2483–2492

Molecular Biology and Evolution

Analyzing the Evolution of RNA Secondary Structures in Vertebrate Introns Using Kimura's Model of Compensatory Fitness Interactions

Robert Piskol and Wolfgang Stephan

Department of Biology II, Section of Evolutionary Biology, Ludwig-Maximilians-University, Munich, Germany

Previous studies have shown that splicing efficiency, and thus maturation of pre-mRNA, depends on the correct folding of the RNA molecule into a secondary or higher order structure. When disrupted by a mutation, aberrant folding may result in a lower splicing efficiency. However, the structure can be restored by a second, compensatory mutation. Here, we present a logistic regression approach to analyze the evolutionary dynamics of RNA secondary structures. We apply our approach to a set of computationally predicted RNA secondary structures in vertebrate introns. Our results are consistent with the hypothesis of a negative influence of the physical distance between pairing nucleotides on the occurrence of covariations, as predicted by Kimura's model of compensatory evolution. We also confirm the hypothesis that longer local secondary structure elements (helices) can accommodate a larger number of covariations, wobbles, and mismatches. Furthermore, we find that wobbles and mismatches are more frequent in the middle of a helix, whereas covariations occur preferentially at the helix ends. The GC content is a major determinant of this pattern.

Introduction

After the introduction of the concept of epistatic fitness interactions by Haldane (1931) and Wright (1931), this topic became of great interest to evolutionary geneticists. Originally, epistatic interactions were understood as interactions between genes that are expected to lead to nonrandom associations of polymorphisms between loci (Stephan 1996). However, this effect has only rarely been observed in natural populations. For this reason, we proposed an extension of the epistasis concept to intragenic interactions (Stephan 1996; Chen et al. 1999). Compensatory evolution of RNA secondary structures was suggested as a case of such epistatic selection (Kirby et al. 1995).

RNA structures are comprised of pairing regions and unpaired parts of the RNA sequence. If the structure, and thus the function of an RNA molecule, is more important than its sequence, epistatic selection is expected to retain the form of the structure. Therefore, single mutations within the pairing region of an RNA should be deleterious and selected against, if they destroy the structure. The original conformation, however, can be restored by a mutation that creates a complementary base on the opposite strand. This structure-restoring mutation is called "compensatory." The complete process of reestablishing the pairing by mutations from one canonical base pair to another leads to a "covariation."

If a mutation occurs in introns or at synonymous positions, it is usually assumed to be neutral as it does not change the protein. However, selection is still possible in such regions due to the various stages the mRNA molecule has to pass through during its maturation. Thus, not only the primary sequence of an mRNA or pre-mRNA but also its secondary and tertiary structures may play a role and be subject to selective pressure. Our interest in this study is directed toward introns because they may show coevolutionary patterns that are less confounded by other processes than those of coding regions (i.e., they do not underlie the selective constraints imposed by the coding function of a sequence).

The analysis of the evolution of compensatory mutations can be based on a two-locus, two-allele model described by Stephan (1996). It incorporates Kimura's (1985) idea of compensatory neutral mutations, which states that individual mutations are deleterious but harmless in certain Watson–Crick base pair configurations. Under the assumption of a randomly mating diploid population, 2 linked loci with alleles A, a at locus 1 and B, b at locus 2 are examined. Alleles A and B may mutate to a and b, respectively. However, no back mutations are allowed due to the small probability of multiple hits. On the passage from AB to ab, two different intermediate conformations aB and Ab are possible. Assuming a genic selection scheme, these may have fitnesses $1 - s_1$ and $1 - s_2$, whereas the initial and end states have fitness 1. The recombination fraction between both loci is described by parameter r.

In the context of RNA structures, A and B may be identified with the bases adenine (A) and uracil (U), respectively, whereas a and b are guanine (G) and cytosine (C). The intermediate configurations will then be AC and GU and assigned the selection coefficients s_1 and s_2, respectively (usually $s_1 > s_2 > 0$, as GU pairs are more stable than AC pairs).

The rate of compensatory evolution k_c is defined as the inverse of the expected transition time from state AB to ab. Depending on the strength of selection against deleterious intermediate states, k_c may be influenced by recombination. In the case of strong selection, k_c decreases exponentially with increasing r. This is due to the fact that recombination removes newly established beneficial double mutants (ab) from the gene pool. In the case of weak selection, k_c is independent of r (Stephan 1996). Thus, under strong selection against intermediate states, it is predicted that the number of observed covariations within RNA structures decreases as the distance, and hence the recombination rate, between loci increases—the so-called distance effect (Stephan and Kirby 1993; Stephan 1996; Chen et al. 1999).

With the growing availability of comparative genomics data and sophisticated means for RNA secondary structure prediction, we are able to test the distance effect hypothesis and other properties of RNA secondary structures. To this end, we used computationally predicted RNA secondary structures in vertebrate introns (Pedersen

Key words: RNA secondary structure, covariation, compensatory mutation, introns.

E-mail: piskol@bio.lmu.de.

Mol. Biol. Evol. 25(11):2483–2492. 2008
doi:10.1093/molbev/msn195
Advance Access publication September 4, 2008

et al. 2006). These structures were inferred from sequence alignments of eight species covering a wide phylogenetic range (from humans to teleost fishes).

Materials and Methods

Data Source (RNA Secondary Structures and Alignments)

RNA secondary structures (folds) were downloaded from the hg18.evofold track of the UCSC Table Browser (http://genome.ucsc.edu/) (Karolchik et al. 2004). They were originally predicted by Pedersen et al. (2006) using EvoFold—a comparative genomics method that relies on phylogenetic stochastic context-free grammars. The structures are based on a MULTIZ alignment of genome-wide sequences of four to eight species (Schwartz et al. 2003; Blanchette et al. 2004), including human, chimpanzee, mouse, rat, dog, chicken, zebrafish, and puffer fish genomes. The alignments of these eight species were directly parsed from the web output of the UCSC Genome Browser (Kent et al. 2002). However, the annotation of single and double substitutions as given by the Genome Browser was discarded because it is based on the human sequence as the reference. All sequences are genomic DNA sequences, hence thymine is used instead of uracil. Reference genes and miRNA annotations were obtained from the UCSC Table Browser tracks hg18.refGene and hg18.wgRNA (on 16 April 2008 and 24 April 2008). For each of the folds, the GC content was taken as reported by the Genome Browser for the genomic location of the fold including a region of 500 nt on both sides.

We are interested in evolutionary events that took place at the level of local structural elements (helices) in an RNA fold. We therefore defined a helix as a region in the secondary structure that is encompassed on both sides by interior, hairpin, or multiloops (Zucker and Stiegler 1981). For each of the helices, we annotated the alignment columns that contained covariations and also identified wobble (GT) and mismatch base pairs. Furthermore, we determined the length, the average substitution rate, and the GC content of each helix as well as the average free energy of each fold. The helix length was taken as the number of pairing brackets in the RNA structure of that region. The average substitution rate was obtained by calculating the expected number of substitutions between each pair of sequences in the alignment (Jukes and Cantor 1969), divided by the number of compared positions and averaged over all pairwise comparisons. Here, only positions in the alignment that were not involved in a covariation were included. The GC content of a helix was calculated as the number of GC base pairs in the alignment of all species and divided by the number of all canonical base pairs (GC + AT). The average free energy was calculated using RNAeval (Hofacker et al. 1994) for each of the sequences in the alignment and averaged over all sequences.

Fold Selection Criteria

The data source constitutes a large set of high-quality RNA secondary structures. However, we retained only helices from folds that satisfied the following criteria: 1) folds located in introns, 2) only one splice form per gene, 3) fold length $\geq$50 nt, 4) alignment of $\geq$7 species, 5) helix length $\geq$3 nt, and 6) the fraction of substitutions in all pairs of sequences per helix $\leq$3/4.

These criteria were chosen for the following reasons. Intronic regions were of special interest as they are not under the selective pressure to code for amino acids. Because it is assumed that the fragments under investigation have a certain function (as structures are conserved over a wide phylogenetic range), they are expected to compensate for disruptions in the RNA structure. This makes Kimura's (1985) model of epistatic interactions applicable. The functionality of RNA secondary structure and the applicability of Kimura's model were demonstrated in many systems. For example, it was found that the disruption of a hairpin structure in intron 1 of the alcohol dehydrogenase gene in *Drosophila melanogaster* is associated with a significantly reduced splicing efficiency and protein production (Chen and Stephan 2003). A compensatory mutation, however, restored the original efficiency. We chose only folds with a size of at least 50 nt because in such folds, recombination would have had a sufficiently high chance to act such that a distance effect would be expected. By choosing alignments of seven or eight species, we ensured a sufficiently wide phylogenetic range that allowed us to detect covariations. It also guaranteed a higher alignment quality due to the higher conservation in this region. The substitution rate between pairs of sequences in each helix was used as a hint for correctly aligned sequences. A wrongly aligned sequence might have led to an incorrect prediction of covariations. We therefore removed helices from the data set if any pairwise comparison of sequences in their alignment yielded more than 75% differences. In addition, we removed complete folds if, on average, the sequences showed an exceptionally high free energy when folding them according to the predicted RNA secondary structure with RNAeval. This was taken as a sign for a misalignment of at least one sequence. Sequences with a high free energy do not fit well into the predicted RNA structure based on the complete alignment and the chance for them to appear in nature in that fold is low.

Logistic Regression

To monitor factors that are responsible for the occurrence of covariations or substitution events in general (covariations, wobbles, or mismatches), we applied a logistic regression approach. We assumed that the occurrence of these events is independent of each other (e.g., the occurrence of a covariation at a certain position in the helix does not depend on the occurrence of covariations at other positions of the helix). Under this simplifying assumption, the probability of observing a certain number of covariations in a helix follows a binomial distribution:

$$P(Y = y) = \binom{n}{y} p^y (1 - p)^{n-y}, \quad (1)$$

where n is the number of helix positions (the helix length), y the number of observed substitutions in that helix, and p the probability of observing one of these events. We simplified

this process by noting only the presence or absence of an event at each position of the helix. This reduced the binomial process in equation (1) to a Bernoulli process. Although n and y are known, the probability p of observing a substitution is not known and has to be inferred from the available data. This task can be accomplished by applying a regression approach. In general, regression analysis opts to describe the relation between a dependent variable Y and independent variables $\boldsymbol{x}=(x_1,\ldots,x_i)$ in the least complex way. In the case of covariations, the probability of observing one covariation given several independent variables can be described by

$$p = P(Y = 1|\boldsymbol{x}) = \pi(\boldsymbol{x}), \tag{2}$$

where

$$\pi(\boldsymbol{x}) = \frac{\exp(\beta_0 + \beta\boldsymbol{x})}{1 + \exp(\beta_0 + \beta\boldsymbol{x})} = F(\beta_0 + \beta\boldsymbol{x}). \tag{3}$$

F is the logistic distribution function with coefficients β_0 and $\beta=(\beta_1,\ldots,\beta_i)$. These coefficients are determined by a linear predictor after applying the logit transformation,

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta\boldsymbol{x}. \tag{4}$$

In regression analysis, the insignificance of an independent variable x_k in the model can be understood as conditional independence between response and influence. When modeling data with a binary response, these responses are strictly bounded in the interval [0, 1]. A linear regression model might predict unwanted values that are smaller than 0 or grater than 1, while the logistic curve asymptotes at 0 and 1. Logistic regression can further account for a nonconstant variance, it can be applied to data that is not normally distributed; and in addition, as in the case of covariations, rare events can be modeled (Crawley 2005). Furthermore, $\exp(\beta_k)$ can be interpreted as the multiplicative change in odds of observing an event given a change of variable x_k by one unit (Supplementary Material online).

Logistic regression models can be set up using different numbers of independent variables. The more variables are used, the better the resulting fit. However, retention of some of these variables may not significantly improve the fit. To balance model simplicity and fit, we applied a model simplification procedure whenever necessary. Therefore, first the automatic procedure based on Akaike's information criterion in R (R Development Core Team 2006) was used to determine insignificant variables. Close-to-significant variables that still remained in the model were then tested for their justification by analysis of deviance.

Independent Variables in Logistic Regression Analysis

When using logistic regression to identify forces that may be responsible for the occurrence of covariations, wobbles, or mismatches, one has to decide on variables that describe aspects of the RNA secondary structure. Kimura's (1985) model suggests that at least three parameters are needed to explain the evolution by compensatory mutations: the rate of recombination between pairing nucleotides, the strength of selection against individual mutations, and the mutation rate. In our study, we used the distance in sequence between the two interacting nucleotides in a helix (x_1) instead of the recombination rate. To describe the strength of selection, we used 1) the length of a helix (x_2), 2) the distance of the mutated position to the end of the helix (x_3), and 3) the average substitution rate in the helix (x_4). The latter variable also covers properties of the mutation rate, as does the GC content of a helix that was included in our analysis as variable x_5. Depending on whether the occurrence of covariations or wobbles/mismatches was under investigation, a slightly different measurement of these variables was applied. The most accurate description of these variables can be achieved by recording them for all pairing nucleotides in each sequence of the alignment one by one. The exact distance in sequence between two pairing positions and to the end of the helix were obtained in this way. This approach, however, is only applicable to wobbles and mismatches because their identification does not depend on the phylogenetic tree. In the case of covariations, it was not always discernable which species displayed the ancestral base pair and which one the base pair emerging from the covariation event. Thus, for the analysis, it would have been necessary to omit many covariations. We overcame this problem by sacrificing some accuracy in measuring the influencing variables and included all observations of covariation events. That is, we chose as response variable the presence or absence of a covariation in a column of the complete alignment. It was therefore not necessary to know in which species the substitutions occurred. However, to obtain the distance between two columns, it was necessary to average the distance over all sequences and also to average the distance to the end of the local secondary structure element. The remaining three variables helix length, average substitution rate, and GC content per helix were determined for covariations, wobbles, and mismatches as described in the Data Source (RNA Secondary Structures and Alignments).

To take the difference in the recombination rate between autosomes and X chromosome into consideration, the distances on the X chromosome were multiplied by 2/3. This compensates for the fact that males have only one X chromosome, which reduces the overall possibility of the X chromosome to recombine.

Some of the influencing variables are correlated with each other (supplementary material table S1, Supplementary Material online), which may lead to the effect of multicollinearity in the logistic regression analysis. The strongest correlation is between the position in a helix and its length (Pearson: $\rho = 0.5344$, $P < 2.2 \times 10^{-16}$). Multicollinearity may render the results of the logistic regression hard to interpret. It can lead to insignificant coefficients even though the independent variable is important and can also result in very large confidence intervals (CIs) for the regression coefficients (Belsley et al. 1980). These intervals may contain 0 and make it impossible to tell the direction of the influence variable. The presence of multicollinearity was tested by the

variance inflation factor (VIF) (Chatterjee and Price 1991), which in case of multicollinearity exhibits values higher than 10 (Menard 1995).

Results

Data sets and Identification of Covariations

We applied the selection criteria outlined in Materials and Methods to the complete data set, thus reducing it from 47,510- to 507-folds. These structures represent the first set of data to be analyzed and will be denoted as data set *DS1*. They contain 3,116 local structural elements (helices), of which 246 hold a total of 284 covarying columns in their alignments (i.e., some helices contain more than one covariation). In four of these columns, two covariations are found, respectively, which results in a total of 288 covariations. Yet, our logistic regression approach based on a binary response variable takes only the presence or absence of a covariation in a column into account regardless of the number of covariations at this position.

It is obvious that not all of these helices will underlie the same amount of evolutionary constraint. Some parts of the fold may carry an important role in the function of the complete RNA molecule, whereas others may be of less relevance. The helix that connects the 3′ and 5′ ends of the sequence might have a significant role because it determines the closure of the structure. Therefore, we collected all such helices from *DS1* into a smaller data set *DS2*. In 21 of 507 folds from *DS1*, the closing helix was shorter than 3 nt and is thus not considered here. This resulted in a set of 486 closing helices, of which 40 contained a total of 51 columns with covariations.

These two data sets were analyzed for the evolutionary time point at which a covariation occurred. In *DS1*, only for 94 of the 284 columns could it be unambiguously determined which species contained the ancestral base pair and which the derived one. These covariations are shown in figure 1 at the branches of the tree. The number of covariations is expected to be identical in branches of the same length in the phylogenetic tree. However, the current data set shows a clear overrepresentation of covariations in branches leading to zebrafish and pufferfish. This finding may be attributed to relaxed selective constraints in these species due to whole-genome duplications as described by Christoffels et al. (2004) and Volff (2005) and references therein. Such reduced constraints play an important role in the evolution of duplicated genes (Ohta 1988). They are expected to not only lead to a higher number of compensatory substitutions but also reduce the distance effect (Kimura 1985).

In total, 18 of the 507 fragments in *DS1* overlap with miRNA structures. To check whether these structures have different evolutionary constraints than the rest of the folds, the average substitution rates of the helices in miRNA and non-miRNA folds were compared. Because these rates did not show any significant difference (supplementary table S2, Supplementary Material online), all folds were retained in the data set. We also accounted for putatively different selective pressures between helices by including the average substitution rate into the models.

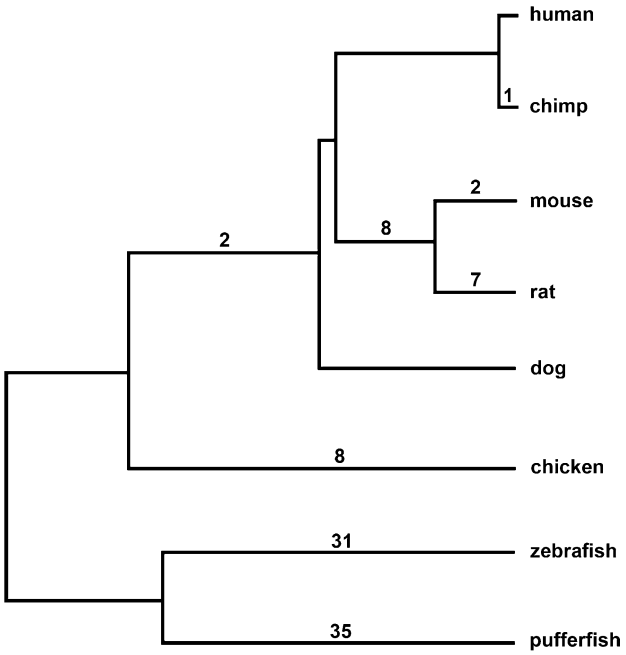

FIG. 1.—Phylogenetic tree of the taxa used in the analysis (Pedersen et al. 2006). Counts at branches give the number of covariations with known covariation direction.

Logistic Regression Models for Covariations

We first investigated factors that influence the occurrence of covariations by setting up logistic regression models for *DS1* and *DS2* (table 1). In this configuration, the response variable corresponds to the presence or absence of a covariation in a column of the alignment, whereas the independent variables are chosen as described in Materials and Methods. Aside from the selection criteria that were applied above, we chose only columns that had an average distance of more than 50 nt. This ensured that recombination acting on these regions was sufficiently frequent. In both data sets, relatively few covariations were found—in *DS1*, 126 columns contain a covariation, whereas in 6,445 columns, covariations are absent (47 vs. 2,596 in *DS2*).

Data set *DS1* shows borderline significance in the negative effect of the distance, whereas in *DS2*, the distance has a highly significant negative influence. This can be attributed to the effect of recombination, which removes newly established double mutants from the population. Its effect is stronger in *DS2*, which may indicate a stronger selection pressure on the helices in this data set. In figure 2 in *DS2*, the fitted probabilities of observing a covariation are plotted against the distance. The solid line resembles predicted probabilities at different distances based on the estimated coefficients in logistic regression and holding all other free variables constant at their mean. Due to the vast overrepresentation of columns that do not contain a covariation, the fitted probabilities are very low. Nonetheless, the negative trend induced by the distance variable is clearly visible.

The distance has a negative effect, while the length of a helix positively affects the occurrence of covariations. Both data sets show a significantly positive estimate.

Table 1
Minimal Logistic Regression Models for Covariations in Data Sets *DS1* and *DS2*

	DS1		*DS2*	
Independent Variable	Estimate	Pr(>|z|)	Estimate	Pr(>|z|)
Intercept	−4.6976	$<2 \times 10^{-16}$***	−4.3285	3.89×10^{-16}***
Distance	−0.0036	0.0540	−0.0107	0.0050**
Helix length	0.0869	0.0019**	0.1167	0.0006***
Distance to helix end	−0.1914	0.0260*	−0.3040	0.0240*
Average substitution rate	—	—	—	—
GC content	2.2326	1.13×10^{-16}***	2.5933	0.0001***

NOTE.—The response variable represents presence or absence of a covariation event in the column of the alignment. Missing values indicate the conditional independence of the response on the respective variable. Significance levels: $*P < 0.05$; $**P < 0.01$; $***P < 0.001$.

The same influence of this variable was also found by Parsch et al. (2000) in *Drosophila* data sets. In shorter helices, the smaller number of covariations per position may be due to stronger constraints because single mutations would lead to different helical structures. These constraints are relaxed in longer helices leading to a higher tolerance for mutations, which may explain the greater number of covariations.

Furthermore, in both data sets, the distance to the end of a helix shows a significantly negative effect. It suggests that covariations have a tendency to appear at the ends of local structural elements.

The average substitution rate, on the other hand, does not play a role in the occurrence of covariations. This is not surprising as we calculated the substitution rate by excluding columns that contained covariations. The GC content influences covariations in a highly significant positive manner. Indeed, its effect is estimated to be the most reliable in the model (smallest p values).

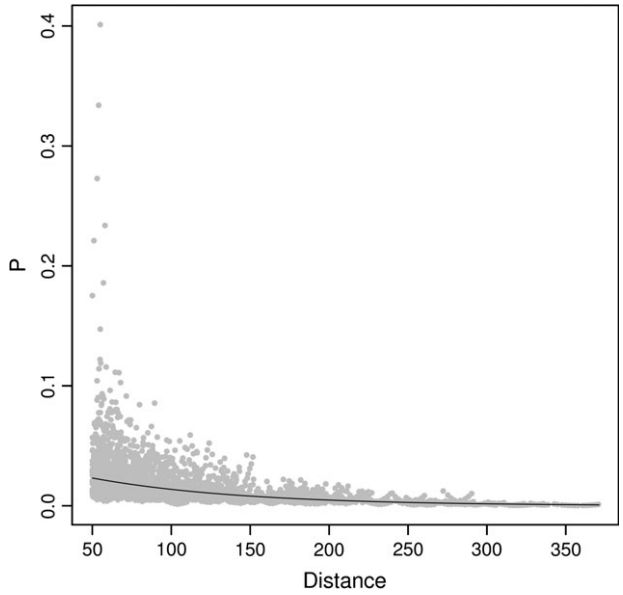

FIG. 2.—Fitted probabilities and predicted probabilities for covariations in *DS2*. Based on the estimated coefficients for covariations in table 1 each gray dot represents the fitted probability (P) for a datapoint of observing a covariation given a certain distance between pairing columns. The solid line represents predicted probabilities as a function of the distance for the estimated coefficients from table 1, holding the remaining variables at their mean.

In general, the estimated coefficients for all independent variables differ widely in their orders of magnitude due to the different scales of the influence variables. The VIF values for all of the coefficients are less than 1.3 (supplementary table S3, Supplementary Material online). Based on the common threshold of 10 for the VIF, this suggests that multicollinearity is not a problem in our analysis. Even if the CIs of each of the coefficients are large (table 2), none of them includes 0. This allows us to infer the direction of the influence variable even though no definitive conclusion about its exact value can be made.

For instance, given the CI for the distance variable in *DS2* (−0.0190 to −0.0040), the increase in distance between two columns by one unit results in a relative change in the odds of observing a covariation by a factor between 0.9812 and 0.9960. Although these numbers represent only a small decrease for the change by one unit, a change by 50 units (nucleotides) would alter the initial odds by a factor between 0.3871 and 0.8184. The distance estimate from table 1 (−0.0107) would yield a change by a factor of 0.5857 ($=\exp(-0.0107)^{50}$). Therefore, the odds of observing a covariation are reduced by nearly one-half when the distance increases by only 50 nt—a rather strong effect.

Logistic Regression Models for Wobbles and Mismatches

We further tested how the influence variables in *DS1* affect the occurrence of wobble and mismatch events. Therefore, the response variable was chosen to indicate the presence or absence of a wobble or mismatch, respectively. Again, all base pairs with a distance smaller than 50 nt were removed from the data. A total of 42,164 canonical base pairs, 4,739 wobble pairs, and 522 mismatch pairs

Table 2
CIs for Logistic Regression Estimates from Table 1

	DS1		*DS2*	
Independent Variable	2.5%	97.5%	2.5%	97.5%
Intercept	−5.3505	−4.0566	−5.4229	−3.2538
Distance	−0.0074	−0.0001	−0.0190	−0.0040
Helix length	0.0286	0.1387	0.0444	0.1810
Distance to helix end	−0.3640	−0.0261	−0.5824	−0.0524
Average substitution rate	—	—	—	—
GC content	1.4646	2.9992	1.2808	3.9132

Table 3
Minimal Logistic Regression Models for Wobbles and Mismatches in *DS1*

	DS1(wobbles)		*DS1*(mismatches)	
Independent Variable	Estimate	Pr(>|z|)	Estimate	Pr(>|z|)
Intercept	−2.6481	$<2 \times 10^{-16}$***	−5.5821	$<2 \times 10^{-16}$***
Distance	−0.0011	9.41×10^{-16}***	—	—
Helix length	0.0558	$<2 \times 10^{-16}$***	0.0453	0.0097**
Distance to helix end	0.0588	2.12×10^{-16}***	0.0828	0.0413*
Average substitution rate	2.6430	$<2 \times 10^{-16}$***	6.9817	$<2 \times 10^{-16}$***
GC content	—	—	1.0391	1.24×10^{-16}***

NOTE.—The response variable represents presence or absence of a wobble/mismatch in a base pair. Missing values indicate the conditional independence of the response on the respective variable. Significance levels are the same as in table 1.

were available for the logistic regression analysis. Table 3 shows the estimated model parameters for wobbles and mismatches after reducing the models as much as possible (removing insignificant influence variables). The distance between pairing nucleotides seems to play a role in the occurrence of wobbles; however, the estimated coefficient and thus the reduction in the odds of observing a wobble is much smaller than for covariations ($\exp(-0.0011)^{50} = 0.9465$). The distance has no effect on mismatches. For both wobbles and mismatches, the helix length plays a significant role, which agrees with previous results for covariations and suggests that wobbles and mismatches tend to occur only in longer helices that have the capacity to tolerate slight disruptions of the structure. Furthermore, both models suggest that wobbles and mismatches are more frequent at greater distances from the helix ends (i.e., in the inner regions of the helix). This is in contrast to the estimates that were obtained for covariations. It shows that suboptimal base pairs may be tolerated in the middle, whereas the ends of helices tend to be preserved by covariation events. Both models agree on a significantly positive effect of the average substitution rate. This was expected as the substitution rate should directly influence divergence. The models, however, differ in the influence of the GC content, which appears to be an important factor only for mismatches. The estimated coefficient is smaller than its counterpart in the model for covariations, indicating that the occurrence of mismatches does not depend as strongly on the GC content. The occurrence of wobbles, on the other hand, does not depend on the GC content at all. Although a higher GC content may lead to a higher mutation rate (Smith et al. 2002; Ochman 2003), the probability of fixation of these mutations appears to be a limiting factor. A possible reason may be that the total number of wobbles in a helix (which on average is much higher than those of mismatches and covariations) is limited to preserve the stability of the helix.

For wobbles and also mismatches, none of the CIs for the coefficients overlap with 0, and the VIF values are also low (<1.49) (supplementary table S4, Supplementary Material online).

Distribution of GC Base Pairs over Helix Positions

In addition to the logistic regression analysis, we investigated several other characteristics of the data sets. These help to validate the reliability of the results and give explanations for effects seen in the logistic regression analysis. One of these factors is the GC content and its distribution over helix positions. In general, intronic regions in *Drosophila* show higher AT levels than neighboring exons (Chen and Stephan 2003). Zhang (1998) found the GC content in human exons to be 53%. If the observation of Chen and Stephan is also valid in humans, then a GC content lower than 53% would be expected. This was found particularly in introns of intermediate and large sizes (Kalari et al. 2006), where the GC content of first introns was as low as ~40%.

The average GC content of all folds in the original data set (Pedersen et al. 2006) is 41.15%. This low value can be attributed to the adjustment of the EvoFold secondary structure prediction algorithm against predictions in GC-rich regions and was verified by a study on data from the ENCODE project (Washietl et al. 2007). For the structures we selected, the average GC content was found to be 37.87% and 38.43% in *DS1* and *DS2*, respectively. These values are even lower than the one obtained for the original data set and suggest that the folds were correctly annotated as intronic.

We were also interested in the distribution of GC nucleotides over helix positions. For functional RNAs, it is believed that the structure of the RNA is more important than its sequence. Hence, helices within a fold should retain their positions within the sequence. Because G and C are bound by three hydrogen bonds, they should be preferably located at helix ends to prevent those ends from breaking apart. On the other hand, as GC base pairs were found to be more mutable than AT base pairs (Smith et al. 2002; Ochman 2003); their distribution over helix positions may also play an important role in the location of compensatory mutations. Indeed, figure 3 shows that the distribution of AT nucleotides (light gray) over helix positions is shifted toward greater values, whereas GC nucleotides (dark gray) occur preferentially at positions 0 and 1. A one-sided Wilcoxon rank-sum test confirmed these results with $P < 2.2 \times 10^{-16}$ and 2.22×10^{-16} for *DS1* and *DS2*, respectively. It supports the initial conjecture that GC pairs are necessary to maintain the stability of the ends of helical regions and thus ensure the form of the fold. Because the above observation is based solely on columns that did not change (columns with covariation events were excluded), the elevated GC content at these positions may account

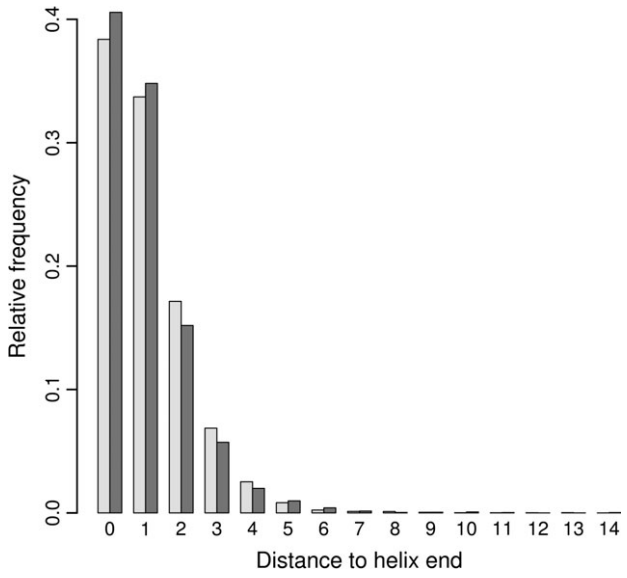

FIG. 3.—Base pair distribution over distances to the nearest helix end in *DS1*. Dark gray bars represent the relative frequencies of GC/CG base pairs, whereas light gray bars give the relative frequencies of AT/TA base pairs at certain distances to the end of a helix.

to some extent for the preferred occurrence of covariations at helix ends.

Distributions of Distance, Helix Length, and Distance to Helix End for Covariations and Wobbles

We also investigated the distributions of the distance, helix length, and the distance to the end of a helix with regard to covariations and wobbles more closely. For a detailed analysis of the distance, data set *DS1* as well as *DS2* were split into two categories depending on the presence or absence of a covariation in a column (fig. 4*a*) and the presence or absence of a wobble in a base pair (fig. 4*b*). Subsequently the distance distributions for both categories were compared using the Wilcoxon rank-sum test. The distances for covarying columns are significantly smaller than distances for columns that do not contain such a substitution event (Wilcoxon one-sided $W = 347{,}685$, $P = 0.0028$ [*DS1*]; $W = 39{,}923$, $P = 2.389 \times 10^{-16}$ [*DS2*]). For wobbles, the distance between wobble base pairs is only significantly smaller in *DS1* ($W = 9{,}645{,}880$, $P = 4.648 \times 10^{-16}$), whereas no significant effect is seen in *DS2* ($W = 15{,}068{,}426$, $P = 0.2040$). These values support the results found in the logistic regression analysis which showed only a very weak effect of distance on wobbles but a more pronounced one on covariations.

In the case of helix length as the variable under investigation, the categories were dependent on the presence or absence of covariations in the helix. The Wilcoxon rank-sum test was then applied to the helix length distributions for both groups (fig. 4*c*). The same was done for helices containing wobbles (fig. 4*d*). Helices containing covariations and wobbles show greater lengths than helices without these substitutions. This observation is significant for wobbles (Wilcoxon one-sided $W = 1{,}696{,}052$, $P < 2.2 \times 10^{-16}$). For covariations, the shift of the distribution containing covariations only exhibits a significant difference in *DS1* (Wilcoxon one-sided $W = 423{,}819.5$, $P = 5.203 \times 10^{-16}$), whereas for *DS2*, it is not significant (Wilcoxon one-sided $W = 9{,}966$, $P = 0.107$).

A similar analysis was applied to the distance to the helix end for covariations and wobbles. For covariations, the two groups consisted of pairing columns containing a covariation and those columns lacking it (fig. 4*e*). In the case of wobbles, base pairs containing a wobble were compared with canonical base pairs (canonical base pairs from all columns that contained a covariation were excluded; fig. 4*f*). The Wilcoxon rank-sum test for *DS1* and *DS2* shows that the distribution of base pairs containing wobbles is shifted toward higher values ($P < 2.2 \times 10^{-16}$ for *DS1* and *DS2*). This corresponds to the results obtained in logistic regression that the influence of the distance to the helix end showed a highly significant positive effect on the occurrence of wobbles. For covariations, the exact opposite effect is found. This is also in accordance with the regression analysis and demonstrates that covariations occur preferably at the ends of helices. A significant result is obtained for *DS1* (Wilcoxon one-sided $W = 2{,}079{,}410$, $P = 0.0038$), whereas the result for *DS2* is only marginally significant due to the small sample size (Wilcoxon one-sided $W = 67{,}353$, $P = 0.0615$). Significance in *DS1* is largely due to columns containing covariations that are located at distance 0 to the helix end (just before an unpaired region). Removing them gives $P = 0.074$.

Base Pairs Involved in Covariations

The compensatory substitutions were also examined for the composition of nucleotide pairs that are involved in such events. We used all 288 covariation events and split the species into groups according to the distinct Watson–Crick base pair configurations for each event (table 4). Because the direction of substitutions was not taken into account, only the upper part of the table is filled. It can be seen that there is an excess of the base pair combinations AT, GC and TA, CG. These base pairs have the ability to mutate into each other via the intermediate step of a wobble pair (AT → GT → GC) by two consecutive transitions. This seems to be the most favorable path as it assures that the disruption of a structure is relatively weak and its functionality is not lost. Because the nucleotides in the wobble intermediate are still associated with each other by two hydrogen bonds, the stability is higher than for noncanonical base pairs. This results in a higher probability to experience a compensatory mutation because the intermediate variant is more frequent in the population. Based on the 94 covariations with known direction, we further observed a slight overrepresentation of covariations that remove GC base pairs from the structure.

Ochman (2003) has shown that the ratio of transitions to transversions in *Escherichia coli* is 2:1. In table 4, we notice an even stronger excess of covariations that arose through two consecutive transitions (the pairs mentioned above) than through two consecutive transversions (all remaining pairs). Here, the beneficial role of a wobble intermediate during the occurrence of a covariation can be seen.

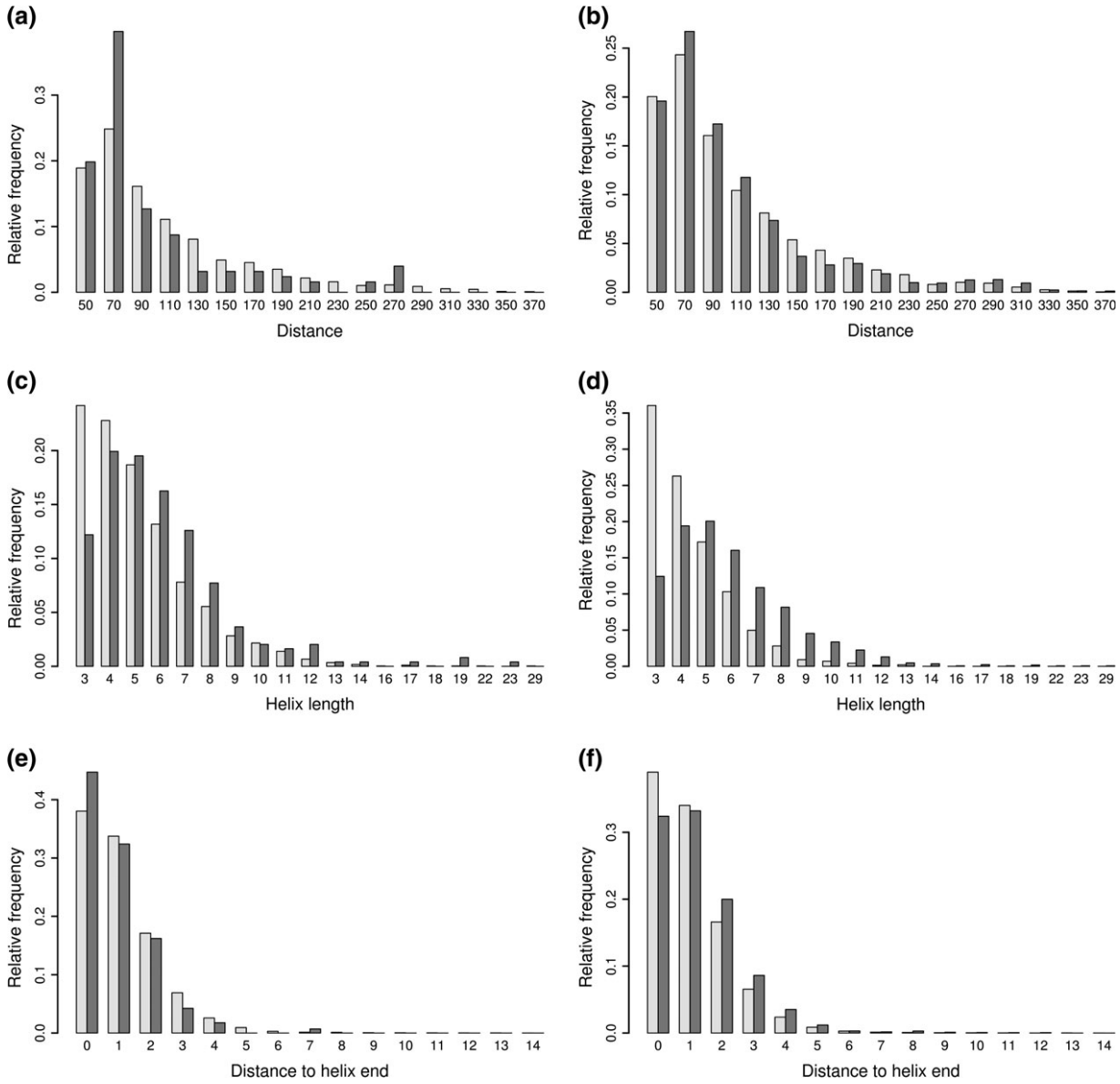

FIG. 4.—Additional statistics for *DS1*. Figures (*a*) and (*b*) show the distribution of distances in *DS1*. Dark bars indicate the distance between columns with covariations (*a*) and the distance between bases that form wobbles (*b*). Numbers on the *x* axis give the midpoint of right-open intervals of length 20. Figures (*c*) and (*d*) show the distribution of helix lengths in *DS1*. Dark bars represent helices containing one or more covariations (*c*) and wobbles (*d*). Figures (*e*) and (*f*) describe the distribution of distances to the helix end for *DS1*. In (*e*) dark bars mark the frequency of occurrences of columns with covariations, whereas in (*f*), they give the frequency of wobble base pairs at a certain distance to the helix end (similarly to the variables used in logistic regression). In all cases, the light gray bars represent the distribution of the respective variable for the class that did not contain the particular substitution event.

Discussion

A preliminary correlation analysis in a *Drosophila* data set (*Drosophila* 12 Genomes Consortium 2007) has suggested that an increasing distance between pairing positions, and hence a higher frequency of recombination, leads to a smaller number of covarying substitutions in RNA structures. Although this study of a very small data set considered the correlation between distance and the number of covariations, our logistic regression analysis presented here has the ability to measure the strength of the influence that comes from the distance variable. Furthermore, and more importantly, the current method has the potential to include additional influence variables into the analysis. Based on Kimura's model of compensatory fitness interactions (Kimura 1985), Parsch et al. (2000) suggested that two main factors are important in the compensatory evolution of RNA structures: the physical distance between pairing nucleotides in a stem region and the length of the stem. These two factors were used in our study as independent variables governing the evolution of covariations. In addition, we included the distance of the mutating position to the nearest end of the helix, the average substitution rate of the helix, and the GC content of the helix.

Table 4
Counts of Base Pair Combinations Involved in Compensatory Substitution Events

	AT	TA	CG	GC	
AT	—	24	14	87	125
TA	—	—	107	33	140
CG	—	—	—	23	23
GC	—	—	—	—	0
					288

NOTE.—The direction of the covariation event is not taken into account. Hence, only the upper part of the table is filled.

By applying logistic regression analysis, we were able to confirm the predicted distance effect for covariations. This effect is particularly strong for helices at the end of folds. According to Kimura's (1985) model, this can be attributed to the circumstance that the distance has an impact on newly established double mutants by making them prone to recombination. Although there is no doubt about the negative influence of distance, great caution has to be taken in inferring the strength of the influence. Due to the large CIs for the distance variable, it is difficult to make quantitative predictions. Although the distance between pairing positions has a negative effect on covariations, the occurrence of wobbles and mismatches seems to be independent of it.

We were also able to confirm that the length of a helix exhibits a positive effect on the occurrence of covariations (Parsch et al. 2000). Furthermore, we found that wobbles and mismatches are more often present in longer helices. Although the majority of helices without substitutions are of length 3, the majority of helices with covariations or wobbles have a length of 4–6 nt (fig. 4*c* and *d*). Relaxed constraints in helices of these lengths seem to play an important role in this observation.

Another important variable we considered is the distance of a nucleotide site to the nearest helix end. Zucker and Sankoff (1984) noticed that wobble base pairs occur frequently in the internal parts of helical structures. We found that this observation holds not only for wobble pairs but also for mismatches (table 3). Because these two classes of changes represent the majority of the variability in the alignment of a helix, we may conclude that lower selection pressure is present in these inner regions. One might expect that this higher variability in inner regions would ensure that covariations are more frequent in the middle of helices. However, the negative estimates for the distance to the helix end in table 1 clearly contradict this and show that covariations tend to be more often present at the ends of a helix, in particular at the outmost position (fig. 4*e*). A possible explanation for this observation may be found in transcription-directed mutagenesis (Burkala et al. 2007). Accordingly, during transcription, the nontranscribed strand is present in a single-stranded form, which makes it susceptible to mutations. The advancing polymerase leads to supercoiling of the single-stranded DNA. Regions that contain inverted complements are thereby arranged into a secondary structure by the pairing of the DNA with itself. Unpaired or mispaired bases in this structure have a mutation rate many times higher than paired ones (Wright 2000) as they are exposed by a greater extent to the soluble environment and thus to nucleotide altering enzymes (Burkala et al. 2007). It was pointed out by Wright et al. (2003) that positions located in close proximity to the stem are the most mutable ones. Therefore, mutations at the ends of helices should be compensated by a second mutation more frequently than at positions that are located within helices. The impact of transcription-directed mutagenesis cannot only be detected experimentally but also leaves signals over evolutionary times (Hoede et al. 2006).

The GC content and its distribution along a helix seems to have a remarkable influence on the observed pattern of variation. The occurrence of covariations and mismatches does generally increase with the GC content. In addition to transcription-directed mutagenesis (described above), it may also contribute to the higher rate of covariations at helix ends as GC bases are more mutable and tend to occur at helix ends. Wobbles, however, are independent of the GC content of a helix, possibly because their high abundance in helices leads to a saturation effect such that new GT mutations can no longer go to fixation.

Finally, we compare our results on the evolution of secondary structures in vertebrates with observations from *Drosophila*. Both influences of distance and helix length were previously analyzed in *Drosophila* data. A study of intronic RNA secondary structures (*Drosophila* 12 Genomes Consortium 2007) that were predicted by RNAalifold (Hofacker et al. 2002) based on six *Drosophila* species found a negative correlation between distance of pairing regions and the rate of compensatory substitutions. Parsch et al. (2000) noticed a positive correlation between helix length and the rate of compensatory evolution in helices of *Adh* introns and *bicoid* 3′ UTR. They also showed a decrease in the rate of compensatory evolution with increasing distance between paired residues, thus confirming our findings. Furthermore, the *bicoid* structure was analyzed by Innan and Stephan (2001). They found the number of substitutions per site in unpaired regions to be higher than in paired regions. Our data shows the same pattern (supplementary table S5, Supplementary Material online).

Although the effects seen in the *Drosophila* data were generally based on small data sets, we rely here on a much larger set of structures. The recent availability of large *Drosophila* data sets from the *Drosophila* 12 Genomes Consortium (2007) makes it now possible to perform the analysis we have done on vertebrate structures on *Drosophila* data as well. Using 125 high-quality RNA secondary structures that were derived from 12 *Drosophila* genomes (Stark et al. 2007), we were able to confirm the negative influence of distance between pairing columns on the rate of compensatory evolution and also the distance to the end of a helix showed the same trend as was found in the vertebrate data.

Supplementary Material

Supplementary material, figures 1 and 3, tables S1–S5, and alignments of the analyzed folds are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank the LMU STABLAB for assistance in the regression analysis and an anonymous reviewer for very valuable comments on the manuscript. This work was supported by grant DFG (Deutsche Forschungsgemeinschaft) Ste 325/8 to W.S.

Literature Cited

Belsley DA, Kuh E, Welsch RE. 1980. Regression diagnostics: identifying influential data and sources of collinearity. New York: John Wiley & Sons, Ltd.

Blanchette M, Kent WJ, Riemer C, et al. (12 co-authors). 2004. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res. 14(4):708–715.

Burkala E, Reimers JM, Schmidt KH, Davis N, Wei P, Wright BE. 2007. Secondary structures as predictors of mutation potential in the lacZ gene of *Escherichia coli*. Microbiology. 153(7):2180–2189.

Chatterjee S, Price B. 1991. Regression analysis by example, 2nd ed. New York: John Wiley & Sons, Ltd.

Chen Y, Carlini DB, Baines JF, Parsch J, Braverman JM, Tanda S, Stephan W. 1999. RNA secondary structure and compensatory evolution. Genes Genet Syst. 74(6):271–286.

Chen Y, Stephan W. 2003. Compensatory evolution of a precursor messenger RNA secondary structure in the *Drosophila melanogaster Adh* gene. Proc Natl Acad Sci USA. 100(20): 11499–11504.

Christoffels A, Koh EGL, Chia J-M, Brenner S, Aparicio S, Venkatesh B. 2004. Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. Mol Biol Evol. 21(6):1146–1151.

Crawley MJ. 2005. Statistics—an introduction using R. New York: John Wiley & Sons.

Drosophila 12 Genomes Consortium. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature. 450:203–219.

Haldane J. 1931. A mathematical theory of natural selection VIII. Metastable populations. Proc Camb Philol Soc. 27: 137–142.

Hoede C, Denamur E, Tenaillon O. 2006. Selection acts on DNA secondary structures to decrease transcriptional mutagenesis. PLoS Genet. 2(11):e176.

Hofacker IL, Fekete M, Stadler PF. 2002. Secondary structure prediction for aligned RNA sequences. J Mol Biol. 319(5): 1059–1066.

Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. 1994. Fast folding and comparison of RNA secondary structures. Monatsh Chem. 125:167–188.

Innan H, Stephan W. 2001. Selection intensity against deleterious mutations in RNA secondary structures and rate of compensatory nucleotide substitutions. Genetics. 159(1):389–399.

Jukes T, Cantor C. 1969. Evolution of protein molecules. In: ed. Munro HN, Mammalian protein metabolism. New York: Academic Press. p. 21–132.

Kalari KR, Casavant M, Bair TB, Keen HL, Comeron JM, Casavant TL, Scheetz TE. 2006. First exons and introns—a survey of GC content and gene structure in the human genome. In Silico Biol. 6(3):237–242.

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. 2004. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 32(Database issue): D493–D496.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. Genome Res. 12(6):996–1006.

Kimura M. 1985. The role of compensatory neutral mutations in molecular evolution. J Genet. 64:7–19.

Kirby DA, Muse SV, Stephan W. 1995. Maintenance of pre-mRNA secondary structure by epistatic selection. Proc Natl Acad Sci USA. 92(20):9047–9051.

Menard S. 1995. Applied logistic regression analysis. Thousands Oaks (CA): SAGE Publications.

Ochman H. 2003. Neutral mutations and neutral substitutions in bacterial genomes. Mol Biol Evol. 20(12):2091–2096.

Ohta T. 1988. Evolution by gene duplication and compensatory advantageous mutations. Genetics. 120(3):841–847.

Parsch J, Braverman JM, Stephan W. 2000. Comparative sequence analysis and patterns of covariation in RNA secondary structures. Genetics. 154(2):909–921.

Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander ES, Kent J, Miller W, Haussler D. 2006. Identification and classification of conserved RNA secondary structures in the human genome. PLoS Comput Biol. 2(4):e33.

R Development Core Team. 2006. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing. Available from http://www.R-project.org.

Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W. 2003. Human-mouse alignments with BLASTZ. Genome Res. 13(1):103–107.

Smith NGC, Webster MT, Ellegren H. 2002. Deterministic mutation rate variation in the human genome. Genome Res. 12(9):1350–1356.

Stark A, Lin MF, Kheradpour P, et al. (46 co-authors). 2007. Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. Nature. 450(7167):219–232.

Stephan W. 1996. The rate of compensatory evolution. Genetics. 144(1):419–426.

Stephan W, Kirby DA. 1993. RNA folding in *Drosophila* shows a distance effect for compensatory fitness interactions. Genetics. 135(1):97–103.

Volff J-N. 2005. Genome evolution and biodiversity in teleost fish. Heredity. 94(3):280–294.

Washietl S, Pedersen JS, Korbel JO, et al. (24 co-authors). 2007. Structured RNAs in the ENCODE selected regions of the human genome. Genome Res. 17(6):852–864.

Wright BE. 2000. A biochemical mechanism for nonrandom mutations and evolution. J Bacteriol. 182(11):2993–3001.

Wright BE, Reschke DK, Schmidt KH, Reimers JM, Knight W. 2003. Predicting mutation frequencies in stem-loop structures of derepressed genes: implications for evolution. Mol Microbiol. 48(2):429–441.

Wright S. 1931. Evolution in Mendelian populations. Genetics. 16(2):97–159.

Zhang MQ. 1998. Statistical features of human exons and their flanking regions. Hum Mol Genet. 7(5):919–932.

Zucker M, Sankoff D. 1984. RNA secondary structures and their prediction. Bull Math Biol. 46:591–621.

Zucker M, Stiegler P. 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res. 9(1):133–148.

Hideki Innan, Associate Editor

Accepted August 26, 2008

Chapter 3

Selective Constraints in Conserved Folded RNAs of Drosophilid and Hominid Genomes

Robert Piskol and Wolfgang Stephan (2011)
Mol. Biol. Evol., 28(4):1519–1529

Molecular Biology and Evolution

Research article

Selective Constraints in Conserved Folded RNAs of Drosophilid and Hominid Genomes

Robert Piskol*,[1] and Wolfgang Stephan[1]

[1]Department of Biology II, Section of Evolutionary Biology, Ludwig-Maximilian-University, Munich, Germany

***Corresponding author:** E-mail: piskol@bio.lmu.de.

Associate editor: Hideki Innan

Abstract

Small noncoding RNAs as well as folded RNA structures in genic regions are crucial for many cellular processes. They are involved in posttranscriptional gene regulation (microRNAs), RNA modification (small nucleolar RNAs), regulation of splicing, correct localization of proteins, and many other processes. In most cases, a distinct secondary structure of the molecule is necessary for its correct function. Hence, selection should act to retain the structure of the molecule, although the underlying sequence is allowed to vary. Here, we present the first genome-wide estimates of selective constraints in folded RNA molecules in the nuclear genomes of drosophilids and hominids. In comparison to putatively neutrally evolving sites, we observe substantially reduced rates of substitutions at paired and unpaired sites of folded molecules. We estimated evolutionary constraints to be in the ranges of $(0.974, 0.991)$ and $(0.895, 1.000)$ for paired nucleotides in drosophilids and hominids, respectively. These values are significantly higher than for constraints at nonsynonymous sites of protein-coding genes in both genera. Nonetheless, valleys of only moderately reduced fitness ($s \approx 10^{-4}$) are sufficient to generate the observed fraction of nucleotide changes that are removed by purifying selection. In addition, a comparison of selective coefficients between drosophilids and hominids revealed significantly higher constraints in drosophilids, which can be attributed to the difference in long-term effective population size between these two groups of species. This difference is particularly apparent at the independently evolving (unpaired) sites.

Key words: RNA secondary structure, selective constraints, selection coefficients.

Introduction

In recent years, it has become obvious that great portions of the genomes of complex organisms are being transcribed to produce noncoding RNAs (ncRNAs) that are involved in a variety of important processes (Amaral et al. 2008). In fact, ncRNAs emerge to be the key players in many developmental systems (Amaral and Mattick 2008) and regulators of diseases (Taft et al. 2010). A secondary structure of the molecule is often necessary to perform its function (MacDonald 1990; Bullock et al. 2003). This structure is composed of double stranded regions (helices) that arise through the formation of Watson–Crick (WC) pairs between complementary nucleotides. If mutations occur in the primary sequence of the molecule, they lead to a disruption of these paired regions, thus changing the conformation of the molecule and impairing or, in the worst case, disabling its original function. Although the original conformation of the molecule in space may be restored through a second (so called "compensatory") mutation at the position opposing the first mutation, intermediate variants will suffer from a selective disadvantage that ultimately will result in reduced evolutionary rates in that region. Knowledge of these rates can further our understanding of constraints imposed on RNA molecules and may reveal the importance of their structures. Various studies focused on the distribution of fitness effects of new mutations (Eyre-Walker et al. 2006; Eyre-Walker and Keightley 2007; Keightley and Eyre-Walker 2007, 2010) and evolutionary constraints in nonprotein-coding DNA of different genera (Halligan et al. 2004; Halligan and Keightley 2006; Előry et al. 2010). Also the process of RNA evolution was studied extensively (Stephan and Kirby 1993; Kirby et al. 1995; Stephan 1996; Chen et al. 1999; Innan and Stephan 2001; Chen and Stephan 2003; Knies et al. 2008; Mimouni et al. 2009). However, an analysis of selective constraints in regions of the nuclear genome that are able to form distinct RNA secondary structures with specific functions has hardly been performed. Previous work on mitochondrially encoded transfer RNAs (mt-tRNAs) in mammals (Meer et al. 2010) suggested that large reductions in fitness have to be expected when mutations in the sequence lead to disruption of the mt-tRNA structure. However, mitochondrial DNA differs from nuclear DNA in several respects (high mutation rates, mode of inheritance, and selection; Parsons et al. 1997; John et al. 2010), which may result in different estimates for evolutionary constraints.

Therefore, this study aims to advance our knowledge of conserved ncRNAs that are encoded in hominid and drosophilid nuclear genomes. The availability of ncRNA data sets that were compiled using the same methodology (Pedersen et al. 2006; Stark et al. 2007) allows us to infer constraints for each of the genera and also permits comparison of constraints between drosophilids and hominids. The main focus of our work is the identification of overall selective constraints in ncRNAs, their variation between different genomic locations, and their differences

between drosophilids and hominids. Furthermore, we aimed to determine the depth of the valleys of reduced fitness that have to be crossed by RNA molecules in the transition from one WC pair to another one and investigated how selective constraints are related to structural features of the ncRNA molecule.

To calculate the detrimental effect of mutations in a sequence of interest, it is necessary to know the rate at which mutations accumulate in neutrally evolving regions (regions that are free of the constraint to perform a certain function). Sites used as neutral standards vary between studies and also between species (Koonin and Wolf 2010). Originally, 4-fold degenerate sites were used as the reference for neutral sequence evolution. However, recently, these positions were also found to be subject to evolutionary constraints and were replaced in their function as a neutral standard by repetitive sequences and intronic sequences in hominids (Előry et al. 2010) and drosophilids (Parsch et al. 2010), respectively. Therefore, we used intronic and intergenic ancestral repeats (ARs) as reference for evolution of ncRNAs in genic and intergenic regions of hominid ncRNAs and positions 8–30 of short introns ($\leqslant$65 nt) as reference for evolution in drosophilid ncRNAs. To calculate the selective pressure against point mutations that disrupt the secondary structure of conserved noncoding RNA molecules and mutations at unpaired positions, we compiled data sets of folded molecules according to distinct genomic regions of Drosophila (focusing on the *Drosophila melanogaster*/*D. simulans* comparison) and vertebrates (human/chimpanzee comparison). We then estimated substitution rates for paired and unpaired nucleotides within the structures as well as for sequences of putatively neutral evolution. Subsequently, we compared the observed numbers of substitutions in the RNA sequences of interest with their expected numbers (which were obtained from the neutrally evolving genomic regions) and determined the selective constraint C and selection coefficient s scaled by the effective population size N_e.

Materials and Methods

Sequence Data

Sequence data for drosophilids and vertebrates were obtained from the University of California Santa Cruz (UCSC) Genome Browser home page (Kent et al. 2002) in form of MULTIZ multiple sequence alignments for *D. melanogaster* (dm3) and human (hg18) genome assemblies. The Drosophila alignment, which consists of up to 12 Drosophila species, was analyzed for the *D. melanogaster* group (*D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta*, and *D. ananassae*). The vertebrate alignment comprised up to 17 species from which we only used eight for our analysis (human, chimpanzee, mouse, rat, dog, chicken, zebrafish, and pufferfish) because these were the basis for the annotation of conserved RNA secondary structure elements by Pedersen et al. (2006). Annotations of functional RNA secondary structures (folds) in Drosophila (Stark et al. 2007) and human (Pedersen et al. 2006) genomes were obtained from the evofold tracks of the Genome Browser and are based on the previously mentioned multiple alignments of 3–12 and 4–8 species, respectively. Scores in the range [0,1] are available for all positions in the folds and describe the confidence of structure annotation. Annotations of genes, repetitive sequences, and nucleotide quality scores were also obtained from the UCSC Genome Browser home page. The position of genes within the *D. melanogaster* and human assembly were taken from the reference gene (refGene) tracks (downloaded on 13 March 2010 and 26 January 2010 for Drosophila and human, respectively). ARs, that is, repetitive sequences common to human and chimpanzee, were determined according to the RepeatMasker annotation (table rmskRM327) for the human assembly. We only considered long terminal repeats, DNA transposons, short interspersed elements, long interspersed elements, and other repeats but excluded simple repeats, low complexity regions, and microsatellites as well as RNA-coding genes from our analysis as described by Előry et al. (2010).

Mapping of Folds

Functional RNA secondary structures were classified based on their location in the genome. We differentiated between nine sequence types. The main distinction was made between sequences in intergenic regions and sequences that overlap with protein-coding genes. Sequences falling into the latter category were further divided according to the following three criteria: 1) number of splice forms (single transcript: ST and alternatively spliced: AS), 2) inclusion into the mature transcript (exon and intron), and 3) translation of region (UTR and non-UTR). The criterion for an intronic location was met if the sequence did not overlap with any exons of alternative splice forms. Hence, sequences that are annotated as intronic are exclusively located in an intron in all splice forms (if there is more than one), whereas sequences annotated as located in an exon either overlap partially with exons, are located completely within an exon, or are (partially) excluded from some of the mature transcripts but present in others. This categorization leads to a total of nine combinations (one intergenic category, eight categories for sequences overlapping with protein-coding genes resulting from the combination of the three aforementioned criteria). Folds on the complementary strand of annotated genes were not taken into account. We chose to investigate ST and AS genes separately due to evidence for higher selective pressures in the latter sequence type (Előry et al. 2010), which may be caused by specific factors for splicing (e.g., exonic splice enhancers and silencers, Parmley et al. 2007 and other highly conserved intronic sequences specific to AS genes, Sorek and Ast 2003). The roles of folded structures also differ substantially depending on their intronic or exonic location. While structured regions in introns predominantly facilitate the correct splicing (Howe and Ares 1997) and efficiency of splicing (Chen and Stephan 2003), folds in coding portions of genes participate in RNA editing (Gott and Emeson 2000; Li et al. 2009) and programmed frameshifting (Farabaugh 1996; Namy et al. 2004).

Furthermore, many folded structures in UTR regions are known to be responsible for correct mRNA localization (Bullock et al. 2003; Irion and Johnston 2007) suggesting the categorization of our data according to this criterion.

Selection Criteria for Folds and Putatively Neutral Sequences

To avoid the wrong estimation of selective constraints due to alignment of nonorthologous sequences or low sequence quality, we applied various selection criteria that were established before (Keightley et al. 2005). We excluded folded regions and ARs from the vertebrate data set if the alignment of human and chimpanzee sequences in that region contained $<$ 50% of aligned nucleotides. We did not allow for more than five mismatches between human and chimpanzee sequences in a window of 25 nucleotides and ten mismatches in a window of 100 nucleotides. In addition, the overall divergence between the two sequences was required to be $<$ 0.1. These criteria for orthology were complemented by requiring $>$ 50% of the chimpanzee sequence to be of high quality (nucleotide quality score $\geqslant$ 40). In addition, folded regions were subjected to criteria similar to those applied in Piskol and Stephan (2008). Thereby, we allowed for the absence of at most one sequence from the alignment of eight vertebrate species, only included folds of low average free energy (as calculated by RNAeval, Hofacker et al. 1994) and discarded folds that overlapped with any kind of repetitive sequences. Apart from removing complete RNA structures and ARs that did not meet the above criteria, we also removed single alignment positions of low quality from fragments that passed the previous checks. These positions were characterized by either a score $<$ 40 in the chimpanzee sequence (applicable to folds and ARs) or a structure confidence score of $<$ 0.9 (only applicable to folds). In case a paired position in a fold had to be removed from the alignment due to its low sequence score, the pairing partner was omitted from the alignment as well. For the Drosophila data set, we relied on the quality of the available sequences, however, we discarded folds and intron sequences from the analysis if two or more species were missing from the alignment of the six species in the *D. melanogaster* group. Short introns in genes with alternative splicing forms or overlapping gene annotation on the same or opposite strand were not considered. We further omitted sequences according to previous divergence estimates for synonymous sites (Cutter 2008). Thereby, we removed folds and introns if divergence between *D. melanogaster* and *D. simulans* was $>$ 0.226. Divergence between *D. ananassae* and any of the other five species was required to be $<$ 1.324. In addition, positions with low confidence ($<$ 0.9) of the base-specific structure annotations were removed from folded regions.

Substitution Rate Estimation

The estimation of substitution rates in paired and unpaired positions of functional RNAs (k_{paired}, k_{unpaired}) and neutrally evolving sequences (k_{neutral}) of the Drosophila and vertebrate data sets was based on the phylogenies for the *D. melanogaster* group (Drosophila 12 Genomes Consortium 2007) and the eight vertebrate species (Pedersen et al. 2006), respectively. Estimations were performed using baseml from the PAML package (Yang 2007). The maximum likelihood results for substitution rates are presented based on the REV + $\Gamma(4)$ + Inv model of sequence evolution (Tavaré 1986), which explains the data best (according to likelihood ratio test: supplementary tables S2 and S3, Supplementary Material online). The estimation of substitution rates for folded regions in each sequence type of drosophilids and hominids, as well as intron regions in drosophilids was performed on a concatenated alignment of all single fragments. Due to the large amount of ARs in the human and chimpanzee genomes, substitution rates for hominid ARs in the nine sequence types were calculated as averages over all 1-Mb windows along the human chromosomes. We did not use dinucleotide substitution models (e.g., RNA7D, Tillier and Collins 1998; RNA16D, Savill et al. 2001) as we were interested in the rate of substitutions at single sites and not in rates for nucleotide pairs. For substitution rates, 95% confidence intervals (CIs) were obtained by bootstrapping 1,000 times by alignment columns (in the case of drosophilid folds and introns as well as hominid folds) or by 1-Mb windows (hominid ARs). When bootstrapping alignments of folded regions by column, we guaranteed the same number of paired and unpaired positions as in the original alignment. The increase of substitution rates through context-dependent substitution processes was taken into account by removing CpG-prone sites from the vertebrate data set (Gaffney and Keightley 2008). Therefore, substitution rate estimates for hominids are provided for non-CpG-prone sites only.

Calculation of Selective Constraints and Selection Coefficients

The calculation of selective constraint for each sequence type was performed according to the two-lineage approach using the formula $C = 1 - \left(\frac{N_{\text{obs}}}{N_{\text{exp}}}\right)$ (Halligan et al. 2004). N_{obs} is the observed number of differences between two sequences of a certain length. N_{exp} denotes the number of expected substitutions in a neutrally evolving genomic region of the same size. Hence, C describes the fraction of mutations that were removed by selection (due to their selective disadvantage). To avoid over/underestimation of constraint due to differences in GC content between the sequence of interest and the neutrally evolving sequences, Halligan et al.'s (2004) method corrects N_{exp} by adjusting the substitution rates of nucleotide changes that alter the GC content according to the equilibrium GC content (GC*). Thereby, we assumed GC* in Drosophila and humans to be 0.37 (Halligan et al. 2004; Duret and Arndt 2008). This correction allows us to compute estimates of C that are free of the GC compositional bias and makes a comparison of C for sequence types of varying GC content possible.

Furthermore, we computed the scaled selection coefficient $N_e s$ using the symmetrical bidirectional model of sequence evolution (Innan and Stephan 2001; eqs. 5a and 6).

Table 1. Nucleotide Composition of Noncoding RNA Folds for Different Sequence Categories in *D. melanogaster* (A) and Human (B).

	Sequence Type	Number of Folds	Number of Sites		GC Content	
			Unpaired	Paired	Unpaired	Paired
A	*D. melanogaster*					
	intron_UTR_ST	454	6,399	9,246	0.5057	0.2582
	intron_UTR_AS	618	7,089	13,372	0.5025	0.2685
	intron_nUTR_ST	1,078	11,788	22,244	0.5169	0.2687
	intron_nUTR_AS	1,233	6,427	23,260	0.5169	0.2702
	exon_nUTR_ST	352	4,636	7,416	0.5611	0.4602
	exon_nUTR_AS	566	7,317	11,998	0.5328	0.4680
	exon_UTR_ST	253	3,661	5,888	0.4665	0.3018
	exon_UTR_AS	374	4,861	8,050	0.4791	0.3454
	intergenic	11,647	68,849	147,618	0.5198	0.2680
	Total/average	16,575	121,027	249,092	0.5113	0.3232
B	*Human*					
	intron_UTR_ST	88	1,056	2,056	0.5473	0.3152
	intron_UTR_AS	147	1,907	4,128	0.5469	0.3023
	intron_nUTR_ST	724	8,223	18,548	0.5449	0.3114
	intron_nUTR_AS	525	6,119	13,764	0.5272	0.3203
	exon_nUTR_ST	3,022	18,914	34,804	0.5861	0.5355
	exon_nUTR_AS	1,719	11,268	20,356	0.5717	0.5310
	exon_UTR_ST	578	5,150	13,440	0.4254	0.3445
	exon_UTR_AS	567	4,917	12,392	0.4434	0.3704
	intergenic	3,042	35,690	78,096	0.5468	0.2955
	Total/average	10,412	93,244	197,584	0.5266	0.3696

NOTE.—Folded molecules in genic regions were grouped into sequence types according to 1) inclusion into the mature transcript (exon, intron), 2) translation of region (UTR, nUTR), and 3) number of splice forms (ST, single transcript; AS, alternatively spliced). The last row for each species gives the total number of folds/sites and the average GC content.

This model is based on Kimura's (1985) idea of compensatory neutral mutations which states that individual mutations are deleterious but harmless in certain combinations. It assumes that the two intermediate (deleterious) allelic states of a compensatory mutation are subject to selective constraints of the same strength $(s_1 = s_2)$ and mutation rates to and from the intermediate states are equal $(\mu_1 = \mu_2)$. The model parameters are effective population size (N_e), mutation rate (μ), and selection coefficient s (the amount by which fitness of the intermediate is reduced). The ratio between the expected waiting time for the appearance of a double mutant that will successfully reach fixation in the population ($T1$) and the expected waiting time for two independent neutral substitutions ($T1_{neu}$) is a function of these parameters. Or, vice versa, after observing the ratio $\frac{N_{exp}}{N_{obs}}$ in the data $\left(\text{which corresponds to } \frac{T1}{T1_{neu}}\right)$, we are able to find a numerical solution for the scaled selective constraint $N_e s$ given that we know the scaled mutation rate $\theta\ (= 4N_e\mu)$. To obtain N_{exp}, it has to be assumed that the two taxa are separated by only a short phylogenetic distance such that no multiple hits have occurred.

Results and Discussion

Composition of Data Sets

We applied the selection criteria outlined in Materials and Methods to the complete set of conserved drosophilid RNA secondary structures, thus reducing it from 22,682 to 16,575 folds. The majority of these folds falls into intergenic regions (11,647), whereas the remaining 4,928 folds are located in positions within protein-coding genes. Among the latter, we distinguished between folds in regions that are annotated as ST genes (2,137 folds) and AS genes (2,791 folds). The position of the RNA secondary structure within the gene was characterized by the inclusion of the region into the mature transcript (intron/exon) and the structure's location in the untranslated or translated portion of the gene (UTR/nUTR), which resulted in four combinations (intron/UTR, intron/nUTR, exon/UTR, and exon/nUTR). From the 47,510 folds in the vertebrate data, we selected 10,412 structures and partitioned them according to the same methodology. Positions of the selected regions are available from the authors upon request. Table 1 shows the number of folds, number of sites, and GC content for each category in *D. melanogaster* and human, respectively. For *D. melanogaster*, the numbers are given for all sites, whereas in humans, only non-CpG-prone sites were considered. The most striking difference between the two data sets is the elevated number of folds in intergenic regions of drosophilids (almost 70.3%). Even though, this sequence type contains the highest percentage of folds in hominids as well (29.1%), their fraction is much lower than in drosophilids and a remarkable part of structures can also be found in coding parts of the genome. The GC content in drosophilids and hominids is significantly lower at paired than unpaired positions. This difference can be attributed to the adjustment of the secondary structure algorithm to favor pairings in AT-rich regions (Pedersen et al. 2006). Furthermore, a clear difference between the GC content in folds and in neutrally evolving sequences (tables 1 and 2) can be observed. The GC content of neutrally evolving positions (table 2) lies between the elevated content at unpaired and reduced

Table 2. Nucleotide Composition of Sequences Used as Neutral Standards.

Taxa	Sequence Type	Number of Sites	GC Content
Drosophilids	Introns	82,678	0.3402
Hominids	Intergenic AR	262,033,297	0.4606
Hominids	Intron AR	35,563,484	0.4848

NOTE.—AR, ancestral repeats.

content at paired sites in both drosophilids and hominids. In hominids, however, the unpaired positions are more similar to neutral sequences in terms of GC content, whereas in drosophilids, the neutral standard matches more closely the GC content of paired nucleotides. In general, this pattern holds for all sequence types except for folds in the coding portion of genes (sequence types: exon_nUTR_ST and exon_nUTR_AS). These contain higher overall numbers of GC nucleotides than the designated neutral standards which is in accordance with the generally higher number of GC nucleotides in coding portions of the genome (Pozzoli et al. 2008). However, nearly all sequences show a nucleotide composition that deviates from the equilibrium GC content of 0.37 in drosophilids and hominids and thus require a correction of constraint estimates as described.

Substitution Rate Variation

Based on the alignments of *D. melanogaster* and *D. simulans* as well as human and chimpanzee, we calculated substitution rates k_{paired}, $k_{unpaired}$ at paired and unpaired sites in conserved noncoding RNAs and k_{neut} for sequences that evolve under putatively neutral conditions (drosophilids: positions 8–30 of short introns, hominids: ARs). Even though further inference of sequence constraints from uncorrected substitution rates should be performed with caution due to differences in GC contents between test and neutral sequences, substitution rates may serve as a first indicator of restrictive conditions for sequence evolution. To account for the impact of a context dependent increase of the mutation rate CpG prone nucleotides were removed from the human/chimpanzee alignment. Figure 1 and supplementary table S1, Supplementary Material online, show substitution rates for the neutral standards as well as paired and unpaired regions of noncoding RNAs of different sequence

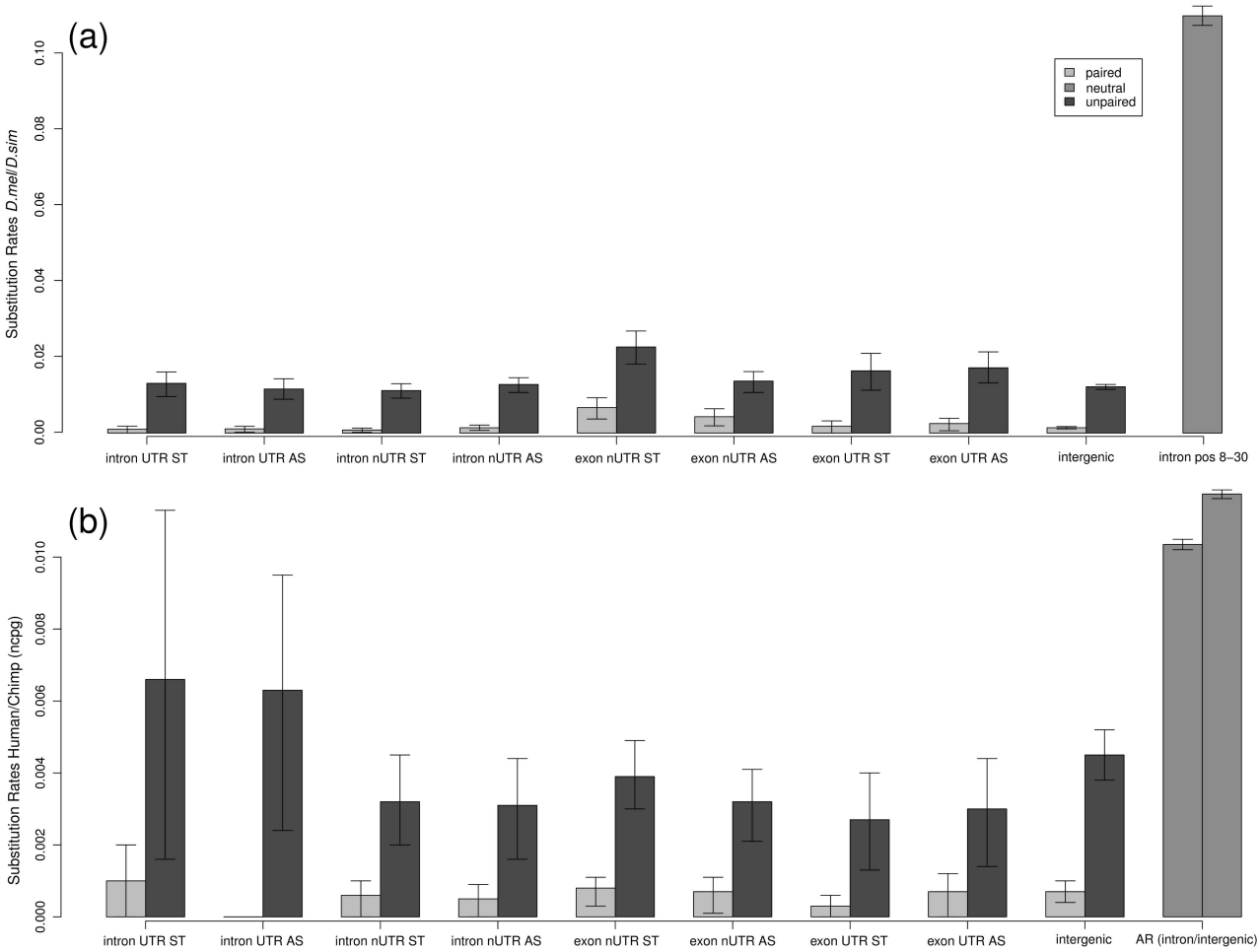

FIG. 1. Substitution rates for folds of various sequence types in the drosophilid (*a*) and hominid (*b*) genomes. Light gray and dark gray bars represent substitution rates in paired (k_{paired}) and unpaired ($k_{unpaired}$) regions, respectively. Substitution rates for neutrally evolving sequences (k_{neut}) are shown in intermediate gray color. Rates for hominids are given at non-CpG prone sites only. Please note the 10-fold difference in the *y*-scale between drosophilid and hominid rates. Sequence types are as in table 1.

type. The values of k_{paired} and $k_{unpaired}$ represent the divergence between *D. melanogaster*/*D. simulans* (fig. 1*a*) and human/chimpanzee (fig. 1*b*), respectively. In comparison to k_{neut}, it is evident that folded RNA molecules evolve under strongly constrained circumstances. This is not only true for paired portions of the folded RNA but also for unpaired regions. Substitution rates are in general lowest at paired positions due to their primary purpose to form and maintain the secondary structure of the RNA molecule ($k_{paired} \ll k_{neut}$) which agrees with the expectation that nucleotide sites evolve slower under dependence (Nasrallah et al. 2011). However, also unpaired sites show considerably reduced rates of molecular evolution ($k_{unpaired} \ll k_{neut}$). Even though they are predicted not to be involved in the formation of the secondary structure, a nucleotide change at an unpaired position may result in a previously not present pairing that leads to a different (for instance energetically) more favorable but functionally defective structure that should be selected against. Furthermore, unpaired sites may still participate in tertiary interactions with other nucleotides of the RNA molecule through canonical WC and other noncanonical base pairs (Leontis and Westhof 2001). Hence, they are also subject to evolutionary constraints and show considerably lower divergence between species than neutrally evolving sites do. While mutations in paired positions inevitably lead to nonisosteric pairs and thus to a strong reduction of the molecule's fitness (if we assume that the fitness is directly associated with the structure), non-WC pairs can be replaced through other isosteric pairs along neutral evolutionary paths (Dutheil et al. 2010). Therefore, selection is stronger on canonical WC pairs that are involved in the formation of the secondary structure than on regions denoted as unpaired, which is reflected by $k_{paired} < k_{unpaired}$.

Variation in divergence between sequence types can also be observed; especially with notably higher values for folds in coding regions of the drosophilid genome. However, these differences may stem from the heterogeneous GC content between sequence types that deviates from GC* to various extents and was previously shown to affect the rates of substitution (Előry et al. 2010). In general, bootstrap CIs are larger for the hominid data and thus suggest greater variation of C within each sequence type along the hominid genome. The obtained estimates for the neutral standard sequences match closely with previous results. Based on 82,678 nucleotides that belong to positions 8–30 of short drosophilid introns, we calculated k_{neut} to be 0.1100 (95% CI = [0.1074, 0.1125]). This value is similar to a recently reported estimate of 0.123 (Parsch et al. 2010). The discrepancy between these two estimates may stem from a different source of alignments and varying sizes of the data sets. Our substitution rate estimates in intronic ARs (0.01035 [0.0102, 0.0105]) and intergenic ARs (0.01175 [0.0116, 0.0119]) in hominids confirm values of a recent study (0.0115 [0.0114, 0.0117]) for the divergence between human and chimpanzee intergenic ARs (Előry et al. 2010). Hence, the drastic difference between k_{paired} and $k_{unpaired}$ in comparison to k_{neut} is not caused by excessively high values of the latter but rather very small substitution rates of the former two. This effect is more pronounced in drosophilids and suggests higher constraints on drosophilid folds than on human folds.

Variation of Selective Constraints

In order to determine differences in selective pressures on noncoding RNA molecules of different sequence type in the drosophilid and hominid genomes, we calculated selective constraints (C) after grouping the molecules according to their genomic location as described in Materials and Methods. Thereby, not only the level of constraint on the pairing portion of the folded structures was of interest but also the degree by which unpaired regions are subject to evolutionary restrictions. Therefore, estimates for constraints in paired (C_{paired}) and unpaired regions ($C_{unpaired}$) were obtained by using positions 8–30 of short introns as a neutral standard for sequence evolution in drosophilids. Intergenic and intronic ARs served as neutral standards for sequence evolution in intergenic and genic hominid regions, respectively. In addition, we obtained estimates for constraint at paired sites by contrasting their evolution with rates at unpaired sites of the structures (C_{paired^*}). Due to considerable restrictions on sequence evolution at unpaired sites (see previous section on substitution rates), we expected to obtain downwardly biased estimates for C_{paired^*}. Nevertheless, it was worthwhile to consider C_{paired^*} for comparison with previous studies (Innan and Stephan 2001) and to observe how constraints relate to each other depending on the use of different neutral standards. Estimates for C_{paired}, C_{paired^*}, and $C_{unpaired}$ (table 3) show that evolution in folded regions is subject to strong constraints. These constraints reach levels higher than at nonsynonymous sites in protein-coding genes and even exceed 0.99 (i.e., $\geqslant$99% of mutations are removed due to purifying selection). From the comparison of paired and unpaired sites, we obtained $C_{paired^*}(\text{dros}) \in (0.681, 0.912)$ for drosophilids and $C_{paired^*}(\text{hom}) \in (0.764, 1.000)$ for hominids. These values approach or even surpass constraints at nonsynonymous sites in these genera ($C_{nonsyn}(\text{dros}) = 0.86$; $C_{0\text{-fold/AS}}(\text{hom}) = 0.759$) (Előry et al. 2010; Parsch et al. 2010). However, unpaired sites are under selective constraints themselves as evidenced from their substantially lower rates of substitution than putatively neutral sites (fig. 1), which suggests that C_{paired^*} is most likely to be underestimated. Indeed, if we use putatively neutral sequences as standard, we observe C_{paired} to be significantly higher than C_{paired^*} (CIs of C_{paired} and C_{paired^*} do not overlap for any category except for folds in UTR regions of ST genes). The values of $C_{paired}(\text{dros}) \in (0.974, 0.991)$ and $C_{paired}(\text{hom}) \in (0.895, 1.000)$ consistently exceed constraints at nonsynonymous sites in the two taxa and suggest that secondary structures are subject to strong evolutionary restrictions. This is also true for unpaired positions as already suggested by the reduced substitution rates. Although our choice to select for conserved regions might lead to increased values of C, it is no exception to also observe such high values in regions that were not chosen according to strong conservation between taxa. For

Table 3. Selective Constraints (*C*) as well as Ratio of Observed to Expected Numbers of Nucleotide Substitutions for Paired (A,B) and Unpaired (C) Sites in folds of Different Sequence Types.

		Drosophilids		Hominids	
	Sequence Type	Constraint *C* (95% CI)	$\frac{N_{exp}}{N_{obs}}$	Constraint *C* (95% CI)	$\frac{N_{exp}}{N_{obs}}$
A					
	intron_UTR_ST	0.989 (0.976, 0.998)	93.246	0.895 (0.682, 1.000)	9.516
	intron_UTR_AS	0.989 (0.980, 0.997)	89.079	1.000 (1.000, 1.000)	NaN
	intron_nUTR_ST	0.991 (0.985, 0.997)	114.900	0.930 (0.861, 0.977)	14.287
	intron_nUTR_AS	0.986 (0.979, 0.992)	73.094	0.937 (0.860, 0.984)	15.977
	exon_nUTR_ST	0.936 (0.911, 0.960)	15.601	0.918 (0.881, 0.955)	12.229
	exon_nUTR_AS	0.974 (0.961, 0.986)	38.113	0.930 (0.876, 0.977)	14.348
	exon_UTR_ST	0.981 (0.962, 0.996)	51.613	0.968 (0.921, 1.000)	31.619
	exon_UTR_AS	0.974 (0.956, 0.988)	39.068	0.932 (0.849, 0.983)	14.791
	Intergenic	0.986 (0.983, 0.988)	69.865	0.937 (0.910, 0.963)	15.765
B					
	intron_UTR_ST	0.910 (0.787, 0.984)	11.099	0.837 (0.231, 1.000)	6.153
	intron_UTR_AS	0.897 (0.805, 0.967)	9.729	1.000 (1.000, 1.000)	NaN
	intron_nUTR_ST	0.912 (0.846, 0.969)	11.368	0.764 (0.488, 0.931)	4.243
	intron_nUTR_AS	0.882 (0.817, 0.935)	8.456	0.795 (0.464, 0.956)	4.874
	exon_nUTR_ST	0.681 (0.532, 0.807)	3.131	0.768 (0.619, 0.873)	4.303
	exon_nUTR_AS	0.783 (0.663, 0.888)	4.598	0.769 (0.553, 0.920)	4.322
	exon_UTR_ST	0.876 (0.741, 0.975)	8.069	0.887 (0.647, 1.000)	8.811
	exon_UTR_AS	0.845 (0.722, 0.933)	6.459	0.787 (0.437, 0.955)	4.688
	Intergenic	0.870 (0.842, 0.894)	7.710	0.836 (0.757, 0.905)	6.102
C					
	intron_UTR_ST	0.878 (0.849, 0.906)	8.190	0.369 (0.000, 0.816)	1.584
	intron_UTR_AS	0.888 (0.864, 0.914)	8.941	0.399 (0.040, 0.703)	1.665
	intron_nUTR_ST	0.894 (0.875, 0.911)	9.409	0.698 (0.578, 0.814)	3.307
	intron_nUTR_AS	0.880 (0.861, 0.897)	8.328	0.700 (0.556, 0.826)	3.336
	exon_nUTR_ST	0.796 (0.758, 0.831)	4.902	0.648 (0.560, 0.734)	2.844
	exon_nUTR_AS	0.876 (0.850, 0.898)	8.051	0.699 (0.599, 0.790)	3.317
	exon_UTR_ST	0.844 (0.799, 0.889)	6.408	0.724 (0.587, 0.860)	3.620
	exon_UTR_AS	0.836 (0.799, 0.875)	6.100	0.693 (0.543, 0.836)	3.253
	Intergenic	0.885 (0.879, 0.891)	8.707	0.631 (0.568, 0.684)	2.712

NOTE.—Values of *C* and $\frac{N_{exp}}{N_{obs}}$ were obtained from the comparisons of (A) paired versus neutral, (B) paired versus unpaired, and (C) unpaired versus neutral sites, respectively. Sequence types are as in table 1.

instance, we observe $C_{paired}(\text{miRNA}[\text{microRNA}]) = 0.68$ and $C_{unpaired}(\text{miRNA}) = 0.41$ as well as $C_{paired}(\text{tRNA}) = 0.83$ and $C_{unpaired}(\text{tRNA}) = 0.76$ for sets of orthologous miRNAs from the mirBase (Griffiths-Jones et al. 2008) and tRNAs from Rfam (Gardner et al. 2009) in human and rhesus macaque, respectively. These values fall into a similar range as our whole-genome estimates even though the phylogenetic range is much smaller. A comparison of constraints in folded molecules to constraints at first, second, and third codon positions that were selected according to the same criteria (supplementary table S4, Supplementary Material online) suggests that RNA molecules are generally subject to larger constraints than protein-coding regions.

We were not able to find significant differences in selective constraints between ncRNAs that belong to different sequence types of the drosophilid and hominid genomes. In drosophilids, we can exclude the low divergence between *D. melanogaster* and *D. simulans* as a potential reason because we did not find any differences when comparing *D. melanogaster* and its more distant relative *D. pseudoobscura* (results not shown). In hominids, the previously visible large variation of substitution rates (previous section) can also be observed when calculating selective constraints and results in large CIs. These observations suggest that evolution of conserved ncRNAs is not influenced by factors that are specific to a certain sequence type (related to a certain genomic context in terms of the location within the gene) but rather by variation in constraints along the genome or intrinsic factors related to the RNA molecule itself.

Previous studies have shown that secondary structures in introns of ST and AS genes are responsible for correct splicing (Eperon et al. 1988) and inclusion or exclusion of exons (Howe and Ares 1997). Furthermore, the importance of other functional elements that are required for alternative splicing was found to result in a larger number of constrained sites in genes with alternative splice forms (Eőry et al. 2010) and might suggest that folded regions in AS genes are under stronger constraints as well. We were, however, not able to observe this pattern which implies that the folded regions in our data set are either 1) not related to splicing, 2) free of this additional constraint, or 3) the constraint on structures related to splicing is as strong as constraints that preserve specific functions in other genomic locations. One notable exception are significantly reduced values of C_{paired} and $C_{unpaired}$ in coding exons

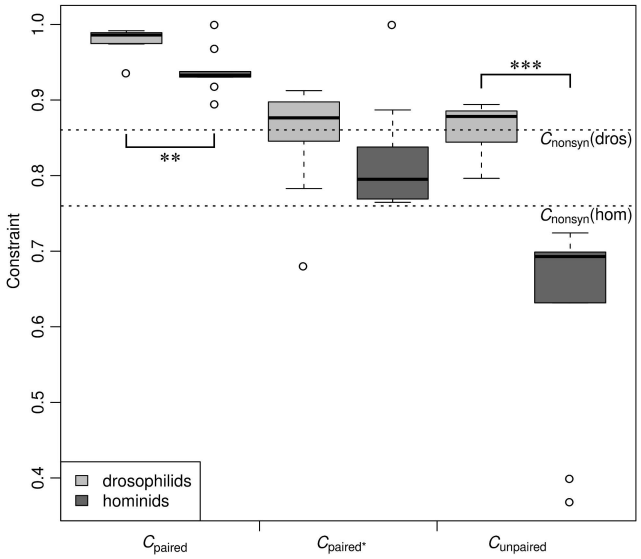

FIG. 2. Selective constraints for drosophilids (light gray) and hominids (dark gray) at paired and unpaired sites. Constraints at paired sites C_{paired} and $C_{paired*}$ were calculated using neutral sites (ARs(hominids)/introns(drosophilids)) and unpaired positions as reference, respectively. $C_{unpaired}$ was calculated from the comparison of unpaired and neutral sites. Brackets indicate significant differences in constraints between drosophilids and hominids (Wilcoxon one-sided test). $^{**}P < 0.01$; $^{***}P < 0.001$. The estimates for constraints at nonsynonymous sites in drosophilids ($C_{nonsyn}(\text{dros}) = 0.86$) and hominids ($C_{nonsyn}(\text{hom}) = 0.76$) are shown as dashed lines. They were taken from Parsch et al. (2010) and Előry et al. (2010), respectively.

of drosophilid ST genes (nonoverlapping CIs for sequence types exon_nUTR_ST and exon_nUTR_AS in table 3).

Even though no significant differences in the strength of selection between sequence regions and between ST and AS genes were detected, there are clear overall differences in selective constraints between drosophilid and hominid folds. As suggested by the theory of nearly neutral evolution (Ohta and Gillespie 1996), there exists a range of selective constraints $\left(\frac{1}{2N_{e(\text{drosophilid})}} < s < \frac{1}{2N_{e(\text{hominid})}}\right)$ that are nearly neutral for a species of small N_e (e.g., hominids) but deleterious for one of large N_e (e.g., drosophilids). Assuming different effective population sizes in the most recent common ancestor (MRCA) of hominids ($N_e \approx 52{,}000$–$96{,}000$; Chen and Li 2001) and drosophilids ($N_e \approx 200{,}000$ *D. melanogaster* worldwide sample; Schug et al. 1998), this fact is then reflected in low constraints in the former and strong constraints in the latter species. Indeed, all estimates for C (except for paired positions in UTR introns of AS genes) follow this theoretical prediction and exhibit lower values for hominids than for drosophilids. When grouping all estimates from different sequence types (table 3), a significant difference between drosophilid and hominid constraints can be observed at paired and unpaired sites (Wilcoxon one-sided $W = 69, P = 0.0053$ (C_{paired}); $W = 81, P = 2.057 \times 10^{-5}$ [$C_{unpaired}$]; fig. 2). It is important to note that the difference for $C_{unpaired}$ between the two genera is much larger than for C_{paired} and $C_{paired*}$. This observation can be explained by a longer fixation time ($\overline{T}$) of a double mutant compared with $\overline{T}$ of a mutation at a single locus for the same $N_e s$ (Kimura 1980, 1985). With growing $N_e s$, $\overline{T}$ (and thus also $\frac{N_{exp}}{N_{obs}}$) increases more slowly at independently evolving sites than for nucleotide pairs (supplementary fig. S1, Supplementary Material online). This in turn results in greater differences of C at independent sites than at paired sites for the same difference in N_e (assuming that s is the same in both genera) and leads to the observed effect in figure 2.

Selection Coefficients Obtained from the Bidirectional Model of Sequence Evolution

Based on the ratio of the expected to observed numbers of substitutions $\left(\frac{N_{exp}}{N_{obs}}\right)$ in paired positions of folded RNA molecules, it is possible to estimate the scaled coefficient of selection ($N_e s$). We assumed $\theta \in (0.003, 0.03)$ for the scaled mutation rate in drosophilids according to previous studies in *D. simulans* and *D. melanogaster* (Kliman et al. 2000; Andolfatto 2001; Hutter et al. 2007) and report further results for the two boundaries of this interval. For humans, $\theta = 0.001$ was chosen according to Nachman and Crowell (2000). Scaled selection coefficients $N_e s$ were obtained from $\frac{N_{exp}}{N_{obs}}$ using the bidirectional model of sequence evolution (Innan and Stephan 2001) for the comparisons of 1) paired with unpaired and 2) paired with putatively neutral sites (fig. 3). For the comparison of paired with unpaired sites, $\frac{N_{exp}}{N_{obs}}$ was taken from table 3 as a value from the interval $(3.131, 11.368)$ and $(4.243, 8.811)$ for drosophilids and hominids, respectively. The same ratio reaches substantially higher values in the range of $(15.601, 114.900)$ for drosophilids and $(9.516, 31.619)$ for hominids when relating paired to putatively neutral sites. These ratios were translated into $N_e s$ using the corresponding values for θ in drosophilids and hominids (fig. 3*a* and *b* and supplementary table S5, Supplementary Material online). From the comparison of paired with unpaired sites, we obtain values for $N_e s$ that fall into the ranges $(0.811, 1.063)$ for hominids, and $(0.735, 1.184)$ and $(0.824, 1.351)$ for drosophilids assuming $\theta = 0.003$ and $\theta = 0.03$, respectively (fig. 3*a*). These estimates for $N_e s$ are only slightly higher than a previously obtained result of $N_e s \approx 0.6$–0.7 from the Drosophila *bicoid* 3'UTR (Innan and Stephan 2001) and suggest that most mutations in paired regions of ncRNAs in hominids and drosophilids are only slightly deleterious $\left(s \lessapprox \frac{1}{2N_e}\right)$. Interestingly, for small θ, relatively low values of $N_e s$ are already sufficient to obtain large $\frac{N_{exp}}{N_{obs}}$ ratios due to the nonlinear (nearly exponential) nature of the relationship between these two parameters. Hence, when relating evolution at paired sites to the neutrally evolving standard in hominids (fig. 3*b*), we obtain $N_e s$ in the range of $(1.088, 1.469)$ and, thus, only slightly higher values for the selection coefficient than from the paired versus unpaired comparison (even though the difference of $\frac{N_{exp}}{N_{obs}}$ between the comparison of paired versus unpaired and paired versus neutral is 2.3- to 3.6-fold). The same is true for estimates in drosophilids: $N_e s \in (1.288, 1.949)$ when θ is assumed to be small

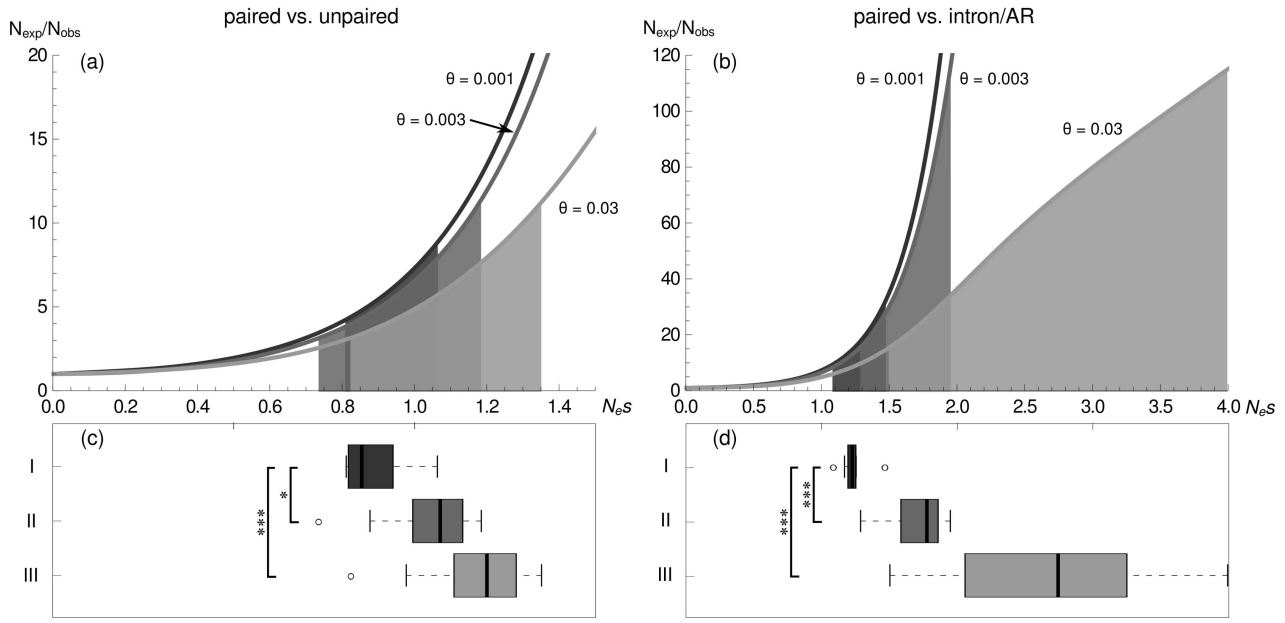

FIG. 3. Scaled selection coefficients at paired sites of folded RNA molecules in hominids ($\theta = 0.001$) and drosophilids ($\theta = 0.003$ and $\theta = 0.03$). Figures 3(*a*) and (*c*) show values obtained from the comparison of paired versus unpaired sites. Figures 3(*b*) and (*d*) show values calculated from the comparison of paired versus putatively neutral sites. Shaded areas under the curves in (*a*) and (*b*) display the range of selection coefficients for the given data. Box plots in (*c*) and (*d*) describe the distribution of selection coefficients in 1) hominids, 2) drosophilids with $\theta = 0.003$, and 3) drosophilids with $\theta = 0.03$. Brackets indicate significant differences in selection coefficients between drosophilids and hominids. $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$.

(i.e., 0.003). If, however, larger values for θ are taken (i.e., 0.03), the estimates for $N_e s$ in drosophilids fall into a broader range (1.504, 3.994) due to a slower increase of $\frac{N_{exp}}{N_{obs}}$ with growing $N_e s$. Although the calculation of C for drosophilids and hominids (previous section) only showed moderate differences at paired positions between the two genera (fig. 2), the comparison of $N_e s$ allows to observe significant differences of the scaled selection coefficient at paired positions (fig. 3*c* and *d*). Again, this difference is due to the different effective population sizes for the MRCA of hominids and drosophilids and leads to nearly neutral effects of mutations in the former but deleterious effects in the latter.

Although compensatory evolution in human mt-tRNAs was shown to cross deep valleys of reduced fitness ($s \approx 10^{-3}$ to 10^{-2}; Meer et al. 2010), our study suggests that evolution in ncRNAs that are positioned in the nuclear genome proceeds through valleys of smaller depth ($s \approx 10^{-4}$). This difference may be attributed to a higher mutation rate in the mitochondrial genome that requires stronger selective pressures that purge deleterious mutations from the sequence.

The estimation of $N_e s$ might be affected by factors that were not accounted for by our analysis. For instance $N_e s$ might be underestimated due to gene conversion that can increase N_{obs} and thus bias the estimates for $N_e s$ downwardly. However, gene conversion alone cannot explain the difference in s by one order of magnitude between folded RNAs encoded on the mitochondrial and nuclear genomes. Furthermore, we evaluated the ratio $\frac{N_{exp}}{N_{obs}}$ of diploid organisms in a haploid framework. It was shown before (Ichinose et al. 2008) that compensatory evolution is accelerated in diploid populations and leads to higher N_{obs}. This is especially true when mutation rate is not low and deleterious effects of mutants are recessive. Again, this suggests that our values of $N_e s$ obtained from the bidirectional haploid model could be underestimated. However, when mutation rate is low ($\theta = 0.01$), which is true in our case, compensatory evolution does not depend strongly on the haploid or diploid selection scheme but the fixation time is rather limited by low mutation rates. Therefore, our use of a haploid model should affect the estimates of $N_e s$ only marginally.

It is important to note that the strength of selection is usually not of a constant magnitude but subject to variation. The values presented here depict the average over all fragments that were used in the analysis—regardless of the underlying structure of the folded molecules. However, we have shown in previous studies that heterogeneity in rates of compensatory evolution is caused by several factors that are related to the two-dimensional structure of the RNA molecule. These include the distance in sequence between pairing nucleotides, the length of pairing regions and the position of the base pair within the pairing region (Parsch et al. 2000; Piskol and Stephan 2008). The influence of these factors may be attributed to the detrimental effect of recombination on double mutations (Stephan and Kirby 1993; Stephan 1996; Chen et al. 1999), the increased tolerance for base pair disrupting mutations, and their influence on stability and structure of the molecule (Mimouni et al. 2009), respectively. This effect of structural variation is not only reflected in different rates of compensatory evolution but can also be seen in variation of selective coefficients (supplementary fig. S2, Supplementary Material online).

Conclusions

The results of our study show the strong restrictions imposed on the evolution of ncRNA molecules. Often the mechanisms these structures are involved in are unknown and hence also the direct source for their constrained evolution. However, we were able to show that their restricted evolution is mostly driven by the basic need of the structure to maintain pairings between nucleotides and just to a smaller extent by the specific region the RNA molecule is located in. It is important to note that the estimation of selective constraints is strongly influenced by the choice of the neutral standard. Our comparison of estimates that were obtained using 1) unpaired regions of folded molecules and 2) repetitive regions and intronic regions shows large differences in sequence constraint C and moderate to strong differences in $N_e s$ depending on the choice of θ. When comparing constraints between drosophilids and hominids, we were able to confirm previous theoretical predictions that species of larger N_e are subject to stronger evolutionary restrictions and that differences in N_e have a greater effect on s at independently evolving sites than at sites that evolve in pairs.

Supplementary Material

Supplementary tables S1–S5 and figures S1–S3 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors would like to thank Olivier Tenaillon and one anonymous reviewer for very valuable comments on the manuscript. This work was supported by grant Deutsche Forschungsgemeinschaft Ste 325/8 to W.S.

References

Amaral PP, Dinger ME, Mercer TR, Mattick JS. 2008. The eukaryotic genome as an RNA machine. *Science* 319(5871):1787–1789.

Amaral PP, Mattick JS. 2008. Noncoding RNA in development. *Mamm Genome*. 19(7–8):454–492.

Andolfatto P. 2001. Contrasting patterns of X-linked and autosomal nucleotide variation in *Drosophila melanogaster* and *Drosophila simulans*. *Mol Biol Evol*. 18(3):279–290.

Bullock SL, Zicha D, Ish-Horowicz D. 2003. The Drosophila hairy RNA localization signal modulates the kinetics of cytoplasmic mRNA transport. *EMBO J*. 22(10):2484–2494.

Chen FC, Li WH. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am J Hum Genet*. 68(2):444–456.

Chen Y, Carlini DB, Baines JF, Parsch J, Braverman JM, Tanda S, Stephan W. 1999. RNA secondary structure and compensatory evolution. *Genes Genet Syst*. 74(6):271–286.

Chen Y, Stephan W. 2003. Compensatory evolution of a precursor messenger RNA secondary structure in the *Drosophila melanogaster Adh* gene. *Proc Natl Acad Sci USA*. 100(20):11499–11504.

Cutter AD. 2008. Divergence times in Caenorhabditis and Drosophila inferred from direct estimates of the neutral mutation rate. *Mol Biol Evol*. 25(4):778–786.

Drosophila 12 Genomes Consortium. 2007. Evolution of genes and genomes on the Drosophila phylogeny. *Nature* 450:203–219.

Duret L, Arndt PF. 2008. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet*. 4(5):e1000071.

Dutheil JY, Jossinet F, Westhof E. 2010. Base pairing constraints drive structural epistasis in ribosomal RNA sequences. *Mol Biol Evol*. 27(8):1868–1876.

Előry 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(1):177–192.

Eperon LP, Graham IR, Griffiths AD, Eperon IC. 1988. Effects of RNA secondary structure on alternative splicing of pre-mRNA: is folding limited to a region behind the transcribing RNA polymerase? *Cell* 54(3):393–401.

Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. *Nat Rev Genet*. 8(8):610–618.

Eyre-Walker A, Woolfit M, Phelps T. 2006. The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics* 173(2):891–900.

Farabaugh PJ. 1996. Programmed translational frameshifting. *Annu Rev Genet*. 30:507–528.

Gaffney DJ, Keightley PD. 2008. Effect of the assignment of ancestral CpG state on the estimation of nucleotide substitution rates in mammals. *BMC Evol Biol*. 8:265.

Gardner PP, Daub J, Tate JG, et al. (11 co-authors) 2009. Rfam: updates to the RNA families database. *Nucleic Acids Res*. 37:D136–D140.

Gott JM, Emeson RB. 2000. Functions and mechanisms of RNA editing. *Annu Rev Genet*. 34:499–531.

Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. 2008. miRBase: tools for microRNA genomics. *Nucleic Acids Res*. 36:D154–D158.

Halligan DL, Eyre-Walker A, Andolfatto P, Keightley PD. 2004. Patterns of evolutionary constraints in intronic and intergenic DNA of Drosophila. *Genome Res*. 14(2):273–279.

Halligan DL, Keightley PD. 2006. Ubiquitous selective constraints in the Drosophila genome revealed by a genome-wide interspecies comparison. *Genome Res*. 16(7):875–884.

Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. 1994. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie* 125:167–188.

Howe KJ, Ares M. 1997. Intron self-complementarity enforces exon inclusion in a yeast pre-mRNA. *Proc Natl Acad Sci U S A*. 94(23):12467–12472.

Hutter S, Li H, Beisswanger S, de Lorenzo D, Stephan W. 2007. Distinctly different sex ratios in African and European populations of *Drosophila melanogaster* inferred from chromosomewide single nucleotide polymorphism data. *Genetics* 177(1):469–480.

Ichinose M, Iizuka M, Kado T, Takefu M. 2008. Compensatory evolution in diploid populations. *Theor Popul Biol*. 74(2):199–207.

Innan H, Stephan W. 2001. Selection intensity against deleterious mutations in RNA secondary structures and rate of compensatory nucleotide substitutions. *Genetics* 159(1):389–399.

Irion U, Johnston DS. 2007. *bicoid* RNA localization requires specific binding of an endosomal sorting complex. *Nature* 445(7127):554–558.

John JCS, Facucho-Oliveira J, Jiang Y, Kelly R, Salah R. 2010. Mitochondrial DNA transmission, replication and inheritance: a journey from the gamete through the embryo and into offspring and embryonic stem cells. *Hum Reprod Update*. 16(5):488–509.

Keightley PD, Eyre-Walker A. 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* 177(4):2251–2261.

Keightley PD, Eyre-Walker A. 2010. What can we learn about the distribution of fitness effects of new mutations from DNA sequence data? *Philos. Trans. R. Soc. B*. 365(1544):1187–1193.

Keightley PD, Lercher MJ, Eyre-Walker A. 2005. Evidence for widespread degradation of gene control regions in hominid genomes. *PLoS Biol*. 3(2):e42.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res*. 12(6):996–1006.

Kimura M. 1980. Average time until fixation of a mutant allele in a finite population under continued mutation pressure: studies by analytical, numerical, and pseudo-sampling methods. *Proc Natl Acad Sci U S A*. 77(1):522–526.

Kimura M. 1985. The role of compensatory neutral mutations in molecular evolution. *J Genet*. 64:7–19.

Kirby DA, Muse SV, Stephan W. 1995. Maintenance of pre-mRNA secondary structure by epistatic selection. *Proc Natl Acad Sci U S A*. 92(20):9047–9051.

Kliman RM, Andolfatto P, Coyne JA, Depaulis F, Kreitman M, Berry AJ, McCarter J, Wakeley J, Hey J. 2000. The population genetics of the origin and divergence of the *Drosophila simulans* complex species. *Genetics* 156(4):1913–1931.

Knies JL, Dang KK, Vision TJ, Hoffman NG, Swanstrom R, Burch CL. 2008. Compensatory evolution in RNA secondary structures increases substitution rate variation among sites. *Mol Biol Evol*. 25(8):1778–1787.

Koonin EV, Wolf YI. 2010. Constraints and plasticity in genome and molecular-phenome evolution. *Nat Rev Genet*. 11(7):487–498.

Leontis NB, Westhof E. 2001. Geometric nomenclature and classification of RNA base pairs. *RNA*. 7(4):499–512.

Li JB, Levanon EY, Yoon JK, Aach J, Xie B, Leproust E, Zhang K, Gao Y, Church GM. 2009. Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science* 324(5931):1210–1213.

MacDonald PM. 1990. *bicoid* mRNA localization signal: phylogenetic conservation of function and RNA secondary structure. *Development* 110(1):161–171.

Meer MV, Kondrashov AS, Artzy-Randrup Y, Kondrashov FA. 2010. Compensatory evolution in mitochondrial tRNAs navigates valleys of low fitness. *Nature* 464(7286):279–282.

Mimouni NK, Lyngsø RB, Griffiths-Jones S, Hein J. 2009. An analysis of structural influences on selection in RNA genes. *Mol Biol Evol*. 26(1)209–216.

Nachman MW, Crowell SL. 2000. Estimate of the mutation rate per nucleotide in humans. *Genetics* 156(1):297–304.

Namy O, Rousset JP, Napthine S, Brierley I. 2004. Reprogrammed genetic decoding in cellular gene expression. *Mol Cell*. 13(2):157–168.

Nasrallah CA, Mathews DH, Huelsenbeck JP. 2011. Quantifying the impact of dependent evolution among sites in phylogenetic inference. *Syst Biol*. 60(1):60–73.

Ohta T, Gillespie JH. 1996. Development of neutral and nearly neutral theories. *Theor Popul Biol*. 49(2):128–142.

Parmley JL, Urrutia AO, Potrzebowski L, Kaessmann H, Hurst LD. 2007. Splicing and the evolution of proteins in mammals. *PLoS Biol*. 5(2):e14.

Parsch J, Braverman JM, Stephan W. 2000. Comparative sequence analysis and patterns of covariation in RNA secondary structures. *Genetics* 154(2):909–921.

Parsch J, Novozhilov S, Saminadin-Peter SS, Wong KM, Andolfatto P. 2010. On the utility of short intron sequences as a reference for the detection of positive and negative selection in Drosophila. *Mol Biol Evol*. 27(6):1226–1234.

Parsons TJ, Muniec DS, Sullivan K, et al. (11 co-authors) 1997. A high observed substitution rate in the human mitochondrial DNA control region. *Nat Genet*. 15(4):363–368.

Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander ES, Kent J, Miller W, Haussler D. 2006. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput Biol*. 2(4):e33.

Piskol R, Stephan W. 2008. Analyzing the evolution of RNA secondary structures in vertebrate introns using Kimura's model of compensatory fitness interactions. *Mol Biol Evol*. 25(11):2483–2492.

Pozzoli U, Menozzi G, Fumagalli M, Cereda M, Comi GP, Cagliani R, Bresolin N, Sironi M. 2008. Both selective and neutral processes drive GC content evolution in the human genome. *BMC Evol Biol*. 8:99.

Savill NJ, Hoyle DC, Higgs PG. 2001. RNA sequence evolution with secondary structure constraints: comparison of substitution rate models using maximum-likelihood methods. *Genetics* 157(1):399–411.

Schug MD, Hutter CM, Wetterstrand KA, Gaudette MS, Mackay TF, Aquadro CF. 1998. The mutation rates of di-, tri- and tetranucleotide repeats in *Drosophila melanogaster*. *Mol Biol Evol*. 15(12):1751–1760.

Sorek R, Ast G. 2003. Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res*. 13(7):1631–1637.

Stark A, Lin MF, Kheradpour P, et al. (46 co-authors) 2007. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature* 450(7167):219–232.

Stephan W. 1996. The rate of compensatory evolution. *Genetics* 144(1):419–426.

Stephan W, Kirby DA. 1993. RNA folding in Drosophila shows a distance effect for compensatory fitness interactions. *Genetics* 135(1):97–103.

Taft RJ, Pang KC, Mercer TR, Dinger M, Mattick JS. 2010. Non-coding RNAs: regulators of disease. *J Pathol*. 220(2):126–139.

Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. In: Miura RM, editor. Some Mathematical Questions in Biology: DNA Sequence Analysis (Lectures on Mathematics in the Life Sciences). Vol. 17. Providence (RI): American Mathematical Society. p. 57–86.

Tillier ER, Collins RA. 1998. High apparent rate of simultaneous compensatory base-pair substitutions in ribosomal RNA. *Genetics* 148(4):1993–2002.

Yang Z. 2007. Paml 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 24(8):1586–1591.

Chapter 4

The role of the effective population size in compensatory evolution

Robert Piskol and Wolfgang Stephan (2011)
Genome Biol. Evol., (accepted with minor revision)

The impact of the effective population size (N_e) on the efficacy of selection has been the focus of many theoretical and empirical studies over the recent years. Yet, the effect of N_e on evolution under epistatic fitness interactions is not well understood. In this study we compare selective constraints at independently evolving (unpaired) and coevolving (paired) sites in orthologous tRNA molecules for vertebrate and drosophilid species pairs of different N_e. We show that patterns of nucleotide variation for the two classes of sites are explained well by Kimura's one- and two-locus models of sequence evolution under mutational pressure. We find that constraints in orthologous tRNAs increase with increasing N_e of the investigated species pair. Thereby, the effect of N_e on the efficacy of selection is stronger at unpaired sites than at paired sites. Furthermore, we identify a "core" set of tRNAs with high structural similarity to tRNAs from all major kingdoms of life and a "peripheral" set with lower similarity. We observe that tRNAs in the former set are subject to higher constraints and less prone to the effect of N_e, whereas constraints in tRNAs of the latter set show a large influence of N_e. Finally, we are able to demonstrate that constraints are relaxed in X-linked drosophilid tRNAs compared to autosomal tRNAs and suggest that N_e is responsible for this difference. The observed effects of N_e are consistent with the hypothesis that evolution of most tRNAs is governed by slightly to moderately deleterious mutations (i.e., $|N_e s| \leq 5$).

4.1 Introduction

The effective population size (N_e) is a fundamental quantity in population genetics. It is essential in shaping neutral nucleotide variation in a population and crucial for determining the efficacy of selection (Kimura, 1983; Charlesworth, 2009). The rate of molecular evolution may decrease, remain unchanged, or increase with increasing N_e, depending on whether mutations are deleterious, (nearly) neutral, or beneficial in nature, respectively (Gillespie, 1999). For independently evolving sites, the rate depends on the product of N_e and the selection coefficient s as well as the scaled mutation rate ($\theta = 4N_e\mu$). Therefore, a mutation that is slightly deleterious in a species of large N_e might have a neutral effect in a species with small N_e (Chamary et al., 2006). This role of N_e in the evolution of independently evolving sites has been studied extensively from a theoretical point of view (Kimura and Ohta, 1969; Ohta, 1972; Kimura, 1983) and has been empirically confirmed (Weinreich and Rand, 2000; Woolfit and Bromham, 2003, 2005; Eőry et al., 2010; Andolfatto et al., 2011). However, the relation between the speeds of evolution due to N_e at independent nucleotide sites and positions that evolve under epistasis is much less clear.

To study the evolution of sites that are involved in epistatic interactions a model with at least two loci is needed. Kimura (1985) introduced a two-locus model of compensatory neutral mutations in molecular evolution. He assumed that mutations at a pair of loci may be individually deleterious but neutral in certain combinations. Given two loci with wild type alleles A and B at the first and second locus, respectively, he studied the expected fixation time ($\overline{T}_{\text{coev}}$) for the double mutant ab under the assumptions that selection against individual mutants is strong (and thus the mutation process is nearly irreversible). Specifically, he assumed that the intermediate configurations of alleles (Ab, aB) suffer the same disadvantage s, and that the wildtype AB and double-mutant ab have the same fitness (i.e., the process does not lead to adaptation but only compensation). Under such conditions ab may rise to fixation without prior fixation of any of the deleterious intermediates (stochastic tunneling) (Iwasa et al., 2004). Subsequently, Kimura's model was extended by incorporating different reductions in fitness (s_1, s_2) for the intermediates Ab, aB (Stephan, 1996) and also allowing for weak purifying selection such that back-mutations may be possible (Innan and Stephan, 2001). In this case fixation of ab can be preceded by a fixation of any of the deleterious intermediates (Ohta, 1973). Fixation times in the two-locus case were also investigated in diploid populations (Ichinose et al., 2008) and for double mutations that lead to adaptation (Lynch, 2010; Weissman et al., 2010). All these models have in common

that for most parameter combinations $\overline{T}_{\text{coev}}$ was found to increase with increasing $N_e s$. Furthermore, for weak selection faster fixation at independently evolving sites is expected, whereas it was shown that in the case of strong selection against deleterious intermediates evolution proceeds faster at coevolving sites (Figure 4.1) (Kimura, 1985).

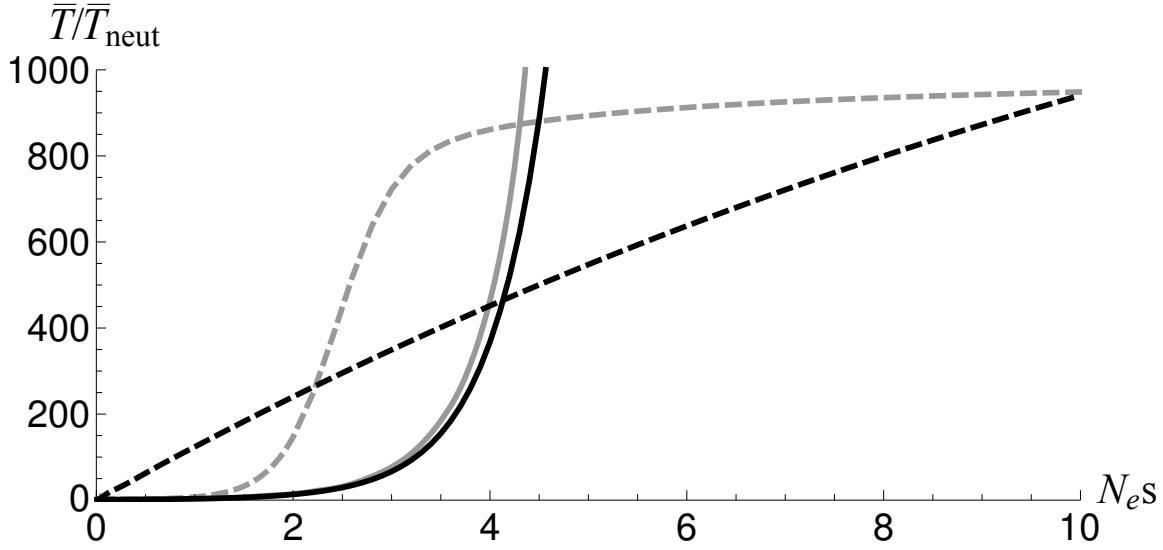

FIGURE 4.1. **Expected ratio of waiting times until fixation of deleterious ($\overline{T}$) and selectively neutral mutations ($\overline{T}_{\text{neut}}$) at independently evolving (solid lines) and coevolving sites (dashed lines).** Black lines describe fixation times in Kimura's unidirectional models (Eq. 13 from (Kimura, 1980) and Eq. 16 from (Kimura, 1985)). Gray lines were obtained by taking back mutations into account, using Eq. 5a and 6 in ref (Innan and Stephan, 2001) for coevolving sites and simulations of the Wright-Fisher process for independent sites. Results are given for a mutation rate $\mu = 2.5 \times 10^{-8}$ and selection coefficient $s = 10^{-4}$.

The role of compensatory mutations has been investigated in the case of protein evolution (Brown et al., 2010), but also RNA molecules provide a great opportunity to directly compare evolution at independently evolving and coevolving sites as they are composed of unpaired nucleotides and nucleotides that form Watson-Crick (WC) base pairs. Previous studies have shown that compensatory mutations are the main driving force of evolution in paired regions of RNA molecules (Chen et al., 1999; Chen and Stephan, 2003; Meer et al., 2010) and that the rate of compensation depends on structural features of the molecule. Specifically, this rate can be related to the length of the pairing region (helix), the position of the pairing nucleotide within the helix, and the GC content of the helix (Parsch et al., 2000; Piskol and Stephan, 2008). Furthermore, population genetic parameters such as the recombination rate between pairing sites were shown to influence the rate of coevolution in RNA molecules (Kirby et al., 1995). Here we investigate how another population genetic

parameter, N_e, shapes RNA evolution. We are especially interested how it influences the rate of evolution at independently evolving and coevolving sites. Therefore, we focus on transfer RNAs (tRNAs) – a class of noncoding RNAs with well studied structure and function. We present a rigorous analysis of selective constraints in tRNA molecules with particular focus on the difference between selective constraints for paired and unpaired nucleotides and interpret the results in the light of theoretical predictions for fixation times of deleterious mutations. In our analysis, the range of moderate and weak purifying selection ($|N_e s| \leq 5$) is of particular interest as evolution of paired sites in noncoding RNA molecules was shown to take place in this parameter range (Piskol and Stephan, 2011).

4.2 Materials and Methods

4.2.1 Sequence Data

Sequence data were obtained from the University of California Santa Cruz (UCSC) Genome Browser FTP server (Kent et al., 2002) in form of axt pairwise alignments for the following vertebrate species pairs: human/macaque (hg19/rheMac2), macaque/marmoset (rheMac2/calJac3), dog/cat (canFam2/felCat3), and chicken/zebra finch (galGal3/taeGut1). The assemblies of these genomes are the same as used by Rfam (Gardner et al., 2009) for the annotation of noncoding RNA families. The pairwise genomic alignment of mouse/rat available at UCSC is based on different genome assemblies than the Celera assemblies (Mural et al., 2002) used by Rfam. Therefore, the Celera assemblies of the mouse and rat genomes were aligned following the same protocol that was used to produce the UCSC alignments. The vertebrate alignments served as a source for orthologous tRNAs and neutrally evolving sequences. Annotations of tRNAs were downloaded from Rfam (Release 10.0) for human, macaque, mouse, rat, dog, and chicken. The UCSC Drosophila multiple alignment, which consists of up to 12 species, was analyzed for the species pairs *Drosophila melanogaster*/*D.simulans* and *D.melanogaster*/*D.yakuba*. It was used to determine neutrally evolving regions only. The annotations of orthologous Drosophila tRNAs were taken from Rogers et al. (2010) and corresponding sequences were downloaded in batch from Flybase (Tweedie et al., 2009). Annotations of protein coding genes were acquired from the refGene tracks of the UCSC Genome Browser for all species except mouse and rat. The locations of ancestral repeats (ARs), i.e., repetitive sequences common to both species in a pair were determined according to RepeatMasker annotations available in the 'rmsk' tables of the UCSC Genome Browser (downloaded on Dec 18,

2010). Protein coding gene annotations in mouse and rat were obtained from Genbank and repeats in the Celera mouse and rat assemblies were annotated using RepeatMasker 3.2.9 (Smit et al., 2010) based on mouse/rat-specific repeat libraries RM-20090604 (Jurka et al., 2005).

4.2.2 Effective Population Sizes

Estimates of long-term effective population sizes were obtained from the literature (Table 4.1) for chicken/zebra finch (Jennings and Edwards, 2005), mouse/rat (Baines and Harr, 2007), and Drosophila (Li and Stephan, 2006). N_e for macaque/marmoset and dog/cat were taken from Piganeau and Eyre-Walker (2009) assuming that the ratio $N_{\text{e-autosomes}} : N_{\text{e-mitochondria}}$ is 4:1. In most of these studies long-term N_e for the pairs were calculated as averages of single-species N_e, which were obtained from polymorphism data. Because no estimate of N_e existed for the pair human/macaque we averaged over N_e for the two species (Eyre-Walker et al., 2002; Evans et al., 2010). However, due to the heterogeneity of the data sources employed for the calculation of N_e, the absolute values were not directly used in our analysis. Estimates of N_e merely served to establish the following semi-quantitative relationship between species pairs: N_e(human/macaque) $< N_e$(macaque/marmoset) $< N_e$(dog/cat) $< N_e$(chicken/zebra finch) $< N_e$(mouse/rat) $< N_e$(*D. melanogaster*/*D. yakuba*) $\approx N_e$ (*D. melanogaster*/*D. simulans*).

4.2.3 tRNA Alignments and Structures

Orthologous vertebrate tRNA sequences and structures for all species pairs were determined based on the pairwise species alignments and Rfam annotations. Thereby, if tRNA Rfam annotations existed for both species in a pair, overlapping orthologs were identified and the corresponding sequences extracted from the pairwise alignment. If Rfam annotations existed only for the reference species, then sequences of the other species ('query' species) that were aligned to the reference in the annotated regions were scored against the tRNA covariance model using cmsearch from the INFERNAL package (version 1.0.2) (Nawrocki et al., 2009). Only hits with an e-value < 0.01 were retained for further analysis. Furthermore, we discarded cases where the query sequence aligned to more than one location in the reference genome and only considered cases where both aligned tRNA annotations were located either on the X chromosome or autosomes in the two species. Subsequently, each pair of orthologous sequences was realigned using cmalign (Nawrocki et al., 2009). To rule out the influence of alignment and structure

prediction on observed selective constraints and to avoid problems with the alignment of unpaired regions we also created alternative alignments using mlocarna (Will et al., 2007) for a structure based alignment and a combination of muscle (Edgar, 2004) and RNAalifold (Hofacker et al., 2002) where alignment and structure are determined separately from each other. Both, mlocarna and RNAalifold rely on thermodynamic predictions of the secondary structure. In some cases thermodynamic prediction may fail to determine the correct topology of tRNA molecules. Therefore, we informed mlocarna and RNAalifold by providing the cmalign structures as constraints for either both sequences or the reference sequence, respectively. Orthologous drosophilid tRNAs from (Rogers et al., 2010) were scored with cmsearch and subsequently aligned using the same three methods as for vertebrate tRNAs. For all pairwise vertebrate and Drosophila alignments only tRNA annotations with an INFERNAL bit score of $S > 35$ in both species were retained. Here, we present results based on the mlocarna alignments. Estimates of selective constraints obtained with muscle and cmalign are shown in Tables C.S2 and C.S3, respectively. They only differ quantitatively, while qualitative predictions are the same for all three methods.

4.2.4 Neutrally Evolving Sequences

Ancestral repeats (ARs) served as indicators for neutral evolution in vertebrates (Előry et al., 2010). Only ARs that reside in intergenic locations were considered. ARs were excluded if the pairwise alignment contained less than 50% of aligned nucleotides. Similar to previous studies (Előry et al., 2010; Piskol and Stephan, 2011), only long terminal repeats, DNA transposons, short interspersed elements, long interspersed elements and other repeats were considered, while simple repeats, low complexity regions and microsatellites were excluded from the analysis. Neutral evolution in drosophilids was based on positions 8–30 in short introns of protein coding genes (Parsch et al., 2010). Thereby only introns of single transcript genes were analyzed to ensure that the sequence is exclusively located in an intron and does not overlap with exons of other splice forms. Introns in genes with overlapping gene annotations on the same or opposite strand were discarded.

4.2.5 Selective Constraints

The strength of selection on a sequence of interest in a species was estimated by calculating the amount of selective constraint C $(= 1 - \frac{N_{\text{obs}}}{N_{\text{neut}}})$, where N_{obs} is the number of observed nucleotide substitutions between two closely related species and N_{neut} is the number of substitutions in a neutrally evolving region of the same length. We obtained N_{obs} in

tRNAs for each species pair by concatenating all single tRNA orthologs. The estimation of constraints may be confounded by several factors. Usually, the rate of substitutions in mammals is increased for dinucleotides in a CpG context through an elevation of the C→G transversion rates after the methylation of cytosine (Siepel and Haussler, 2004). For this reason all CpG-prone sites were excluded from the analysis by removal of all sites that are preceded by a C or followed by a G in the mammalian sequences (Gaffney and Keightley, 2008). Furthermore, it was shown before that the GC content of the sequence and its deviation from the equilibrium GC content (GC*) will lead to increased rates of substitutions (Piganeau et al., 2002; Piskol and Stephan, 2008). Therefore, differences in GC content between species pairs were accounted for by replacing N_{neut} with the expected number of substitutions (N_{exp}) that was calculated from ARs following the method of Halligan et al. (2004). Thereby, substitution rates that change the GC content were adjusted according to GC*, which was assumed to be 0.37 (Halligan et al., 2004; Khelifi et al., 2006; Duret and Arndt, 2008). In all cases, 95% confidence intervals (CIs) for constraints were obtained by bootstrapping the tRNA alignments by column (while ensuring that the number of paired and unpaired columns in the bootstrapped alignment remained the same).

4.3 Results and Discussion

4.3.1 Expected Selective Pressures in RNA Molecules

We used the selective constraint (C) defined by Halligan et al. (2004) as a proxy for the level of selection on tRNA molecules. C describes the portion of deleterious mutations that are removed from the sequence due to purifying selection and is defined as $C = 1 - \frac{N_{\text{obs}}}{N_{\text{exp}}}$ (see Materials and Methods). $\frac{N_{\text{obs}}}{N_{\text{exp}}}$ is equal to $\frac{\overline{T}_{\text{neut}}}{\overline{T}}$ where $\overline{T}$ and $\overline{T}_{\text{neut}}$ are the expected fixation times for deleterious and neutral mutations, respectively (Innan and Stephan, 2001; Piskol and Stephan, 2011). Therefore, the expected values for C can be described in terms of the theoretical predictions for the fixation times as

$$C = 1 - \frac{\overline{T}_{\text{neut}}}{\overline{T}}. \tag{4.1}$$

Due to the dependence of the fixation times on θ and $N_{\text{e}}s$, also C will be influenced by these parameters. The resulting relationship between selective constraints and $N_{\text{e}}s$ (Figure 4.2) for coevolving sites (C_{coev}) and independently evolving sites (C_{ind}) can be obtained by using the expected fixation times ($\overline{T}_{\text{coev}}$ and $\overline{T}_{\text{ind}}$) and their neutral analogs

in equation (4.1), respectively. We used Kimura's unidirectional models for $\overline{T}_{\text{coev}}$ and $\overline{T}_{\text{ind}}$ (Kimura, 1980, 1985) because they are directly comparable in terms of model assumptions and parameters. However, the predictions made here are qualitatively the same as for models that take reversibility of the mutation process into account. Assuming that s is constant between species, the comparison of C_{coev} and C_{ind} allows for three main predictions in the case of weak purifying selection against new mutations:

1. coevolving sites are under stronger selective constraints than independently evolving sites (i.e., $C_{\text{coev}} > C_{\text{ind}}$),
2. constraints increase with increasing effective population size (i.e., $C_{\text{ind}}(N_{\text{e1}}) < C_{\text{ind}}(N_{\text{e2}})$ and $C_{\text{coev}}(N_{\text{e1}}) < C_{\text{coev}}(N_{\text{e2}})$ for $N_{\text{e1}} < N_{\text{e2}}$), and
3. there exists a range of $N_{\text{e}}s$ in which N_{e} has a stronger effect on the evolution at independently evolving than on coevolving sites (i.e., $|C_{\text{ind}}(N_{\text{e1}}) - C_{\text{ind}}(N_{\text{e2}})| > |C_{\text{coev}}(N_{\text{e1}}) - C_{\text{coev}}(N_{\text{e2}})|$).

These general observations are independent of differences in scaled mutation rates (Figure C.S1) and also imply that a change in N_{e} will result in small differences between C for large $N_{\text{e}}s$, but in large differences if $N_{\text{e}}s$ is small (Figure 4.2).

We can use tRNA molecules to test these predictions by assuming that tRNA positions which are not involved in secondary structure formation (here denoted as "unpaired" positions) evolve under the independent model, while changes at nucleotide positions that are involved in WC pair formation with other partners within the sequence ("paired" positions) will be subject to coevolutionary dynamics. It is important for the analysis that N_{e} differs between tRNA molecules. This can be achieved by comparing orthologous tRNAs between species pairs of different long-term N_{e}, but can also be tested within species pairs through the comparison of constraints between X chromosomal and autosomal tRNAs that differ in N_{e}.

4.3.2 Data Set

To investigate the effect of N_{e} on selective constraints in tRNAs we collected data sets of orthologous tRNAs in 7 species pairs of different N_{e} (Table 4.1). We were able to extract approximately the same numbers of orthologous tRNAs for all pairs (a list of genomic positions is available from the authors upon request). Only for murids and birds a smaller number of tRNAs was available. While it might be expected that the amount of identifiable orthologous tRNAs will decrease with increasing divergence between species, we did not observe such a correlation. However, for all species only relatively small numbers of

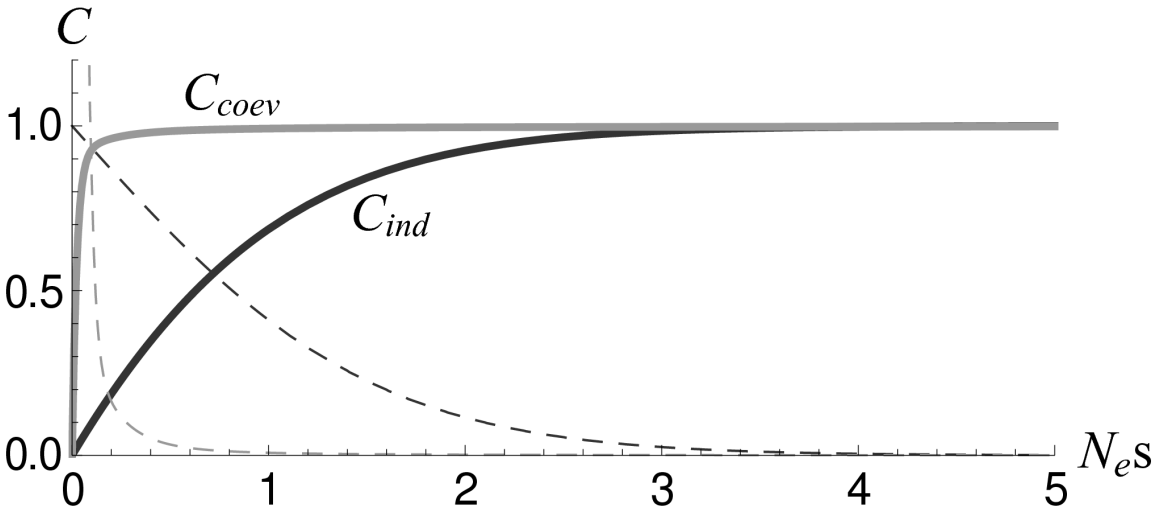

FIGURE 4.2. **Expected selective constraints at independently evolving sites ($C_{\mathbf{ind}}$) and coevolving sites ($C_{\mathbf{coev}}$) as a function of the scaled selection coefficient $N_{\mathbf{e}}s$.** Dashed lines indicate the corresponding slopes. There exists a range of $N_{\mathrm{e}}s$ in which C_{ind} increases more rapidly than C_{coev}. Therefore, the steeper slope for C_{ind} results in a larger difference in constraints at independently evolving sites than at coevolving sites between species with different N_{e}. The trajectories for C_{ind} and C_{coev} were obtained from Kimura's unidirectional models for the expected fixation times of mutant alleles in a population (eq. 13 from (Kimura, 1980) and eq. 16 from (Kimura, 1985)) for a mutation rate $\mu = 2.5 \times 10^{-8}$ and selection coefficient $s = 10^{-4}$.

tRNAs (if any) were identified on the X chromosome compared to the autosomes. This is not due to a low rate of detection of orthologs on the X chromosome, but rather due to a significant underrepresentation of tRNA annotations on the X chromosomes. For instance, the initial set of tRNA annotations in the human genome contained 13 annotations on the X chromosome but 543 on the autosomes. Considering the contribution of the X chromosome to the complete genetic material, 28 tRNAs would have been expected to be located on the X chromosome and 528 on the autosomes, which constitutes a significant deviation from the observed numbers ($X^2 = 8.265$, $P = 0.004$). The same is true for drosophilid tRNAs. In general the GC content in the paired portion of tRNA molecules is larger than for unpaired nucleotides. In particular paired regions in Drosophila and birds show elevated levels of GC nucleotides. For these species no specific increase in mutations due to CpG dinucleotides was expected. Therefore, we did not apply the procedure of Gaffney and Keightley (2008), which usually removes a large portion of guanines and cytosines from the sequence and resulted in a lower number of G and C nucleotides in vertebrates.

Species Pair	N_e	all tRNAs			peripheral tRNAs			core tRNAs		
		# tRNAs	GC content		# tRNAs	GC content		# tRNAs	GC content	
			paired	unpaired		paired	unpaired		paired	unpaired
human/macaque	8.9×10^4	277 (2)	0.5138	0.3105	151 (2)	0.4976	0.3142	126 (0)	0.5316	0.3059
macaque/marmoset	1.7×10^5	268 (1)	0.5172	0.3165	144 (1)	0.4915	0.3173	124 (0)	0.5441	0.3156
dog/cat	5.2×10^5	259 (0)	0.5256	0.3080	134 (0)	0.5206	0.3124	125 (0)	0.5298	0.3033
chicken/zebra finch	6.5×10^5	114 (1)	0.7029	0.4123	63 (1)	0.7149	0.4169	51 (0)	0.6884	0.4062
mouse/rat	$\approx10^6$	106 (0)	0.5552	0.3074	46 (0)	0.5850	0.3271	60 (0)	0.5356	0.2920
D.mel/D.yak	$>10^6$	277 (21)	0.6963	0.3827	95 (5)	0.6788	0.4019	182 (16)	0.7061	0.3720
D.mel/D.sim	$>10^6$	229 (13)	0.6956	0.3822	83 (2)	0.6770	0.4025	146 (11)	0.7071	0.3700

TABLE 4.1. **Composition of tRNA data sets for different species pairs.** Numbers of X-linked tRNAs are given in parentheses.

4.3.3 Core and Peripheral Sets of tRNAs

The total sets of orthologous tRNAs consisted only of molecules that fit to the Rfam tRNA covariance model with high probability (relative to a null model that assumes no structure), which was reflected in INFERNAL bit scores $S > 35$ (see Materials and Methods). We noticed that the distribution of S for most vertebrate pairs is bimodal with a valley at $S \approx 60$ (Figure C.S2). Therefore, the initial set was separated into two subsets according to this value. tRNAs with very high scores ($S \geq 60$) were denoted as a "core" set, because they share great structural similarity with tRNAs from other species in various kingdoms of life. The second ("peripheral") set consisted of tRNAs with lower similarity to the consensus structure of a tRNA ($35 < S < 60$). This partitioning was performed, because we suspected that tRNAs in the core set are under stronger selective constraints, while constraints in peripheral tRNAs are more relaxed. We assumed that under these circumstances N_e will have stronger influence in the peripheral set and will result in more pronounced differences between C, as expected for slightly deleterious mutations. Here, our notion of a core set is based on the structural similarity of tRNAs and differs from the definition of Rogers et al. (2010) who defined a core set based on the conservation of tRNAs throughout the Drosophila genus. Unusually high bit scores may have also been caused by a biased nucleotide composition. However, we did not observe any indication that high scores in our data were related to an exceptionally high or low GC content (Figure C.S3).

4.3.4 The Influence of the Effective Population Size on Constraints in Nuclear-Encoded tRNAs

To test the predictions that are based on Kimura's models for sequence evolution at independently evolving and coevolving sites under continued mutation pressure (Kimura, 1980, 1985) we calculated selective constraints at paired (C_{paired}) and unpaired (C_{unpaired})

positions in orthologous tRNAs for all species pairs (Figure 4.3a, Table C.S1). Thus, we related the rate of molecular evolution in tRNAs to evolutionary rates obtained from the corresponding neutral standard (Table C.S4). Depending on the species pair, the obtained values for C_{paired} and C_{unpaired} fall into the ranges of (0.884, 0.996) and (0.698, 0.982), respectively and thus surpass constraints at nonsynonymous sites in protein coding genes of hominids, murids and drosophilids (Eőry et al., 2010; Parsch et al., 2010). For each species pair we were able to observe significantly higher C_{paired} than C_{unpaired} values (CIs do not overlap), as was expected from the comparison of independently evolving and coevolving sites under Kimura's models. The larger C_{paired} can be explained by the requirement for paired nucleotides to continuously maintain their conformation and thus to preserve the secondary structure of the molecule.

Further examination of Figure 4.3a, in which species pairs were arranged by increasing N_{e} from left to right, indicates that constraints also increase in the same order and verifies that C_{paired} and C_{unpaired} increase with increasing N_{e} – the second prediction that followed from Kimura's models. For instance, the species pairs human/macaque, chicken/zebra finch and *D. melanogaster/D. yakuba*, in this order, have significantly increasing N_{e} in the ranges of 10^4, 10^5 and 10^6, respectively. At the same time the corresponding values of C_{paired} increase from (0.839, 0.933) to (0.942, 0.966) and (0.994, 0.999) and thus significantly differ as well. The same relationship also exists at unpaired sites to an even larger extent. This observation immediately results in the third prediction of Kimura's models, which stated that constraints at independently evolving sites are affected by changes in N_{e} to a larger extent than at coevolving sites. As a result, larger differences can be observed in constraints at unpaired sites between species of different N_{e} than at paired sites. For example the difference in C_{unpaired} between primates and murids $\Delta C_{\text{unpaired}}(prim/mur) = 0.248$, while at paired positions $\Delta C_{\text{paired}}(prim/mur) = 0.105$ and thus much smaller. The same is true for most comparisons between other species pairs. However, as was expected, with increasing N_{e} (and thus also increasing C) the discrepancies between constraints in different species become smaller. Furthermore, the stronger effect of N_{e} on C_{unpaired} also manifests itself in comparisons within species pairs through a decrease in the difference $|C_{\text{paired}} - C_{\text{unpaired}}|$ with increasing N_{e}. In general, these particular patterns in selective constraints are caused by the interplay of $N_{\text{e}}s$ and the influence of mutations rates (Figure C.S1). Thereby, the relationship between C_{ind} and $N_{\text{e}}s$ is mostly independent of θ, while increased C_{coev} are expected when θ is low. This is particularly apparent in the pair human/macaque, which has the lowest θ value (=0.001) in our analysis.

When comparing constraints between species pairs with different divergence (k) it might

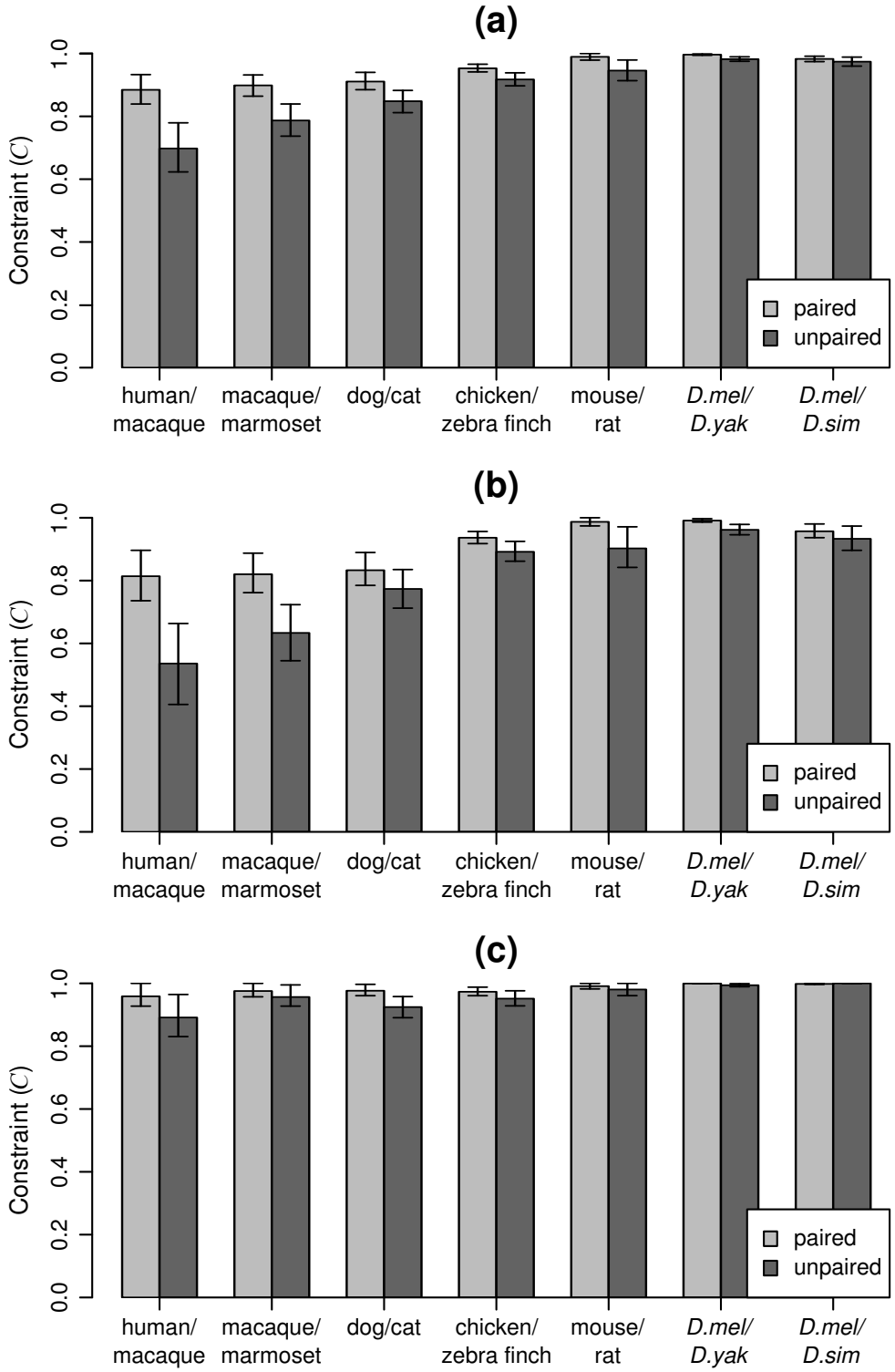

FIGURE 4.3. **Constraint (C) for paired (light gray) and unpaired (dark gray) positions in orthologous tRNAs of different species pairs for (a) the whole data set, (b) peripheral set, and (c) core set.**

have been expected that C increases with increasing k because only orthologous tRNAs with higher conservation can be identified for distant sequences. However, similar to the nonsignificant relation between k and the number of identified tRNAs (n), the relationships between k and C (Kendall's $\tau = -0.43$, $P = 0.24$) as well as n and C (Kendall's $\tau = -0.39$, $P = 0.22$) are not significant. Therefore, we can exclude an influence of divergence and number of identified tRNAs on estimates of constraints in our data. Even if we assume that divergence between chicken/zebra finch and *D. melanogaster*/*D. yakuba* is of a magnitude such that multiple hits cannot be safely ignored (which would result in an underestimation of C for these species), the general pattern persists. For instance, our hypothesis still holds if we replace the *D. melanogaster*/*D. yakuba* pair by *D. melanogaster*/*D. simulans* (which has much smaller divergence and thus lower probability for multiple hits).

4.3.5 Stronger Constraints in Core tRNA Genes

It is also of interest to determine whether the effect of N_e on C is influenced by the overall strength of purifying selection in tRNAs. Therefore, constraints in tRNAs were analyzed after splitting the data into core and peripheral sets. If selection in the former set is strong, then the effect of N_e on C in this set should be low, and vice versa. If, on the other hand, N_e is not responsible for the pattern observed above, the core and peripheral sets should both show signs of approximately equally reduced constraints in species of small N_e. However, the latter assumption can be clearly rejected based on Figures 4.3b,c. Consistent with the assumption that selection pressure is higher in tRNAs belonging to the core set we are able to observe higher constraints in tRNAs from the core set compared to the peripheral set for all species pairs. It is more important, however, that the increase in C with increasing N_e is strong in the peripheral set (Figure 4.3b), while virtually nonexistent in the core set of tRNAs (Figure 4.3c). Therefore, we can assume that selective constraints in tRNAs are most likely influenced by N_e and that this effect is strong if selection is weak, whereas in the case of strong selection our observations follow theoretical predictions, which show that the fixation time of compensatory double mutants is independent of N_e (eq. 8c in ref. (Stephan, 1996)). Even though a separation of the data according to a single score may be crude, our results show that it allows us to distinguish between two sets of tRNAs that seem to be under different selective constraints. Further evidence for this hypothesis comes from the observed GC contents at non-CpG prone nucleotides of the two sets. Compared to the peripheral sets, the core sets show higher GC contents in the paired portion of tRNA molecules for most species

pairs (Table 4.1). A higher GC content was shown to be associated with an increased substitution rate (Előry et al., 2010; Piskol and Stephan, 2011). Therefore, if tRNAs in the core set were subject to the same constraints as tRNAs in the peripheral set, more substitutions would have been expected in tRNAs belonging to the core set. However, the exact opposite is observed, which justifies our separation of the data in two sets and confirms higher constraints in the core set.

4.3.6 Differences in Selective Constraints between Autosomes and X Chromosome

Apart from the differences in N_e between species and their effect on nucleotide variation, effects of N_e on C might also be expected within species. If the contribution of genetic material to the next generation is equal for males and females, the expected ratio of X to autosomal N_e ($\frac{N_{eX}}{N_{eA}}$) is 0.75, due to the presence of the X chromosome in a single copy in males. However, this assumption is not always met. It was reported previously (Hutter et al., 2007) that in an European population of *D.melanogaster* $\frac{N_{eX}}{N_{eA}} = 0.49$ and thus lower than expected, while in an African (ancestral) population of *D.melanogaster* $\frac{N_{eX}}{N_{eA}} = 0.90$. Other studies also suggest that $\frac{N_{eX}}{N_{eA}}$ in ancestral populations may be larger than expected (Andolfatto, 2001; Connallon, 2007; Singh et al., 2007). Therefore, the efficacy of selection may differ between X chromosome and autosomes and may lead to different selective constraints.

To test whether differences in constraints are observed between the X chromosome and autosomes we divided the 277 orthologous tRNAs for the *D. melanogaster*/*D. yakuba* pair according to their genomic location into 21 X-linked and 256 autosomal tRNAs and obtained selective constraints separately for these two sets. It was shown before (Betancourt et al., 2002) that evolutionary rates do not differ between chromosomes in *D. melanogaster*. Nonetheless, we avoided any confounding effects due to systematic differences in mutation rates between X chromosome and autosomes by using introns that were exclusively located on the X chromosome or autosomes as neutral standards for the evolution of X and autosomes, respectively (Table C.S4). Table 4.2A shows constraints in paired and unpaired regions for all X-linked and autosomal tRNAs. Again, paired positions are subject to significantly higher evolutionary constraints than positions that are not involved in the formation of WC base pairs, for both autosomes and the X chromosome. More interestingly, lower constraints can be observed on the X chromosome (C_X) than on the autosomes (C_A) (presumably due to the smaller N_e of the X chromosome). The difference in constraints between autosomes and X

chromosome ($|C_A - C_X|$) is particularly apparent in unpaired portions of tRNAs and is in accordance with theoretical predictions that N_e will have a large impact on evolution at independently evolving sites. Lower constraints on the X chromosome might have also been observed due to reduced evolutionary rates in the neutral standard on the X chromosome rather than increased rates of fixation in tRNAs. However, our neutral divergence estimate for the X chromosome is slightly larger than for autosomes (Table C.S4) and hence cannot be held accountable for lower constraints in X-linked tRNAs, but suggests that a lower C_X at paired and unpaired sites is indeed due to a higher number of fixed differences in tRNAs on the X chromosome. Similar patterns of higher divergence on the X chromosome have also been observed at nonsynonymous sites in the *D. melanogaster* and *D. yakuba* lineages (Begun et al., 2007). Given the estimates of $\frac{N_{eX}}{N_{eA}} > 0.75$ for the ancestral population of *D. melanogaster* from previous studies and assuming that mutations in tRNAs will be mostly slightly deleterious, we would have expected that the rate of fixation on the X chromosome was reduced compared to the autosomes (Vicoso and Charlesworth, 2009; Mank et al., 2010). However, the slightly lower constraints on the X chromosome suggest faster fixations of mildly deleterious mutations in X-linked tRNAs (compared to autosomal tRNAs) and point to a long-term $\frac{N_{eX}}{N_{eA}}$ which is smaller than 0.75 for tRNAs in the *D. melanogaster*/*D. yakuba* pair (see Figure 3 in Vicoso and Charlesworth (2009)).

		C_{paired}	(95% CI)	$C_{unpaired}$	(95% CI)	$\lvert C_{paired} - C_{unpaired} \rvert$
A.	autosomes	0.9977	(0.9961,0.9996)	0.9862	(0.9804,0.9932)	0.0115
	X chromosome	0.9833	(0.9707,1.000)	0.9357	(0.8900,0.9879)	0.0467
	$C_A - C_X$	0.0144	0.031*	0.0505	0.004**	
B.	autosomes	0.9937	(0.9894,0.9989)	0.9698	(0.9547,0.9860)	0.0239
	X chromosome	0.9472	(0.8944,1.000)	0.8369	(0.7073,0.9750)	0.1103
	$C_A - C_X$	0.0455	0.015*	0.1329	0.001**	
C.	autosomes	1.000	(1.000,1.000)	0.9961	(0.9922,1.000)	0.0059
	X chromosome	0.9945	(0.9891,1.000)	0.9744	(0.9488,1.000)	0.0201
	$C_A - C_X$	0.0055	0.085	0.0217	0.067	

TABLE 4.2. **Selective constraints for paired (C_{paired}) and unpaired ($C_{unpaired}$) positions in drosophilid tRNAs located on the autosomes and the X chromosome for (A) the whole data set, (B) peripheral set, and (C) core set.** $C_A - C_X$ is the difference in constraints between tRNAs encoded on the autosomes and X chromosome for paired and unpaired sites. In this case values in the 95% CI column give the p-value for the difference. Significance levels: $^*P < 0.05$; $^{**}P < 0.01$.

In addition, we confirmed that the lower constraints in X-linked tRNAs are in fact significant and did not arise simply by chance due to the small sample size of tRNAs on

the X chromosome. For this reason, we generated 1000 data sets by randomly splitting the 277 Drosophila tRNAs into sets of 21 and 256 instances (resembling the sizes of X and autosomal data). For all repetitions we calculated constraints at paired and unpaired sites in the large and small sets, respectively, and thus obtained distributions for $|C_{\mathrm{A}} - C_{\mathrm{X}}|$ that would be expected at random (Figure 4.4). Indeed, the observed values of $|C_{\mathrm{A}} - C_{\mathrm{X}}|$ are significantly larger than in the randomly assembled sets. This is true for paired regions ($|C_{\mathrm{A}} - C_{\mathrm{X}}| = 0.0144; P = 0.031^{*}$) and to a larger extent in the unpaired portion of tRNAs ($|C_{\mathrm{A}} - C_{\mathrm{X}}| = 0.0505; P = 0.004^{**}$).

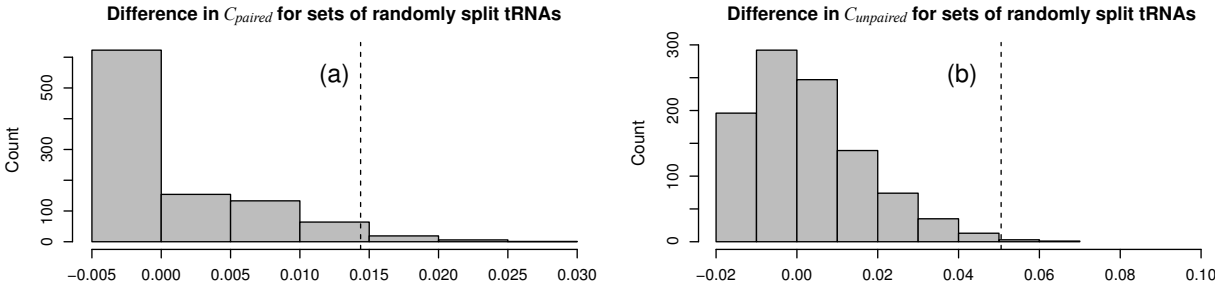

FIGURE 4.4. **Histogram of differences in constraints at (a) paired and (b) unpaired positions between sets of 256 and 21 tRNAs that were created by randomly splitting 277 orthologous tRNAs of *D. melanogaster* and *D. yakuba* 1000 times.** The dashed lines represent the observed values of $|C_{\mathrm{A}} - C_{\mathrm{X}}|$ taken from Table 4.2A.

When repeated separately for tRNAs grouped in core and peripheral sets, the same analysis also supports our previous conjecture that effects of N_{e} on the difference in constraints between X chromosome and autosomes are large if selection is weak (Table 4.2B) but much smaller when selection on the tRNA molecule is overall strong (Table 4.2C). This becomes apparent through significant values of $|C_{\mathrm{A}} - C_{\mathrm{X}}|$ in the peripheral set, while no significant differences are observed in the core set of tRNAs (Table 4.2 and Figures C.S4 and C.S5).

4.4 Conclusions

We showed that divergence patterns in nuclear encoded tRNA molecules of vertebrate and drosophilid species follow general theoretical predictions for sequence evolution under mutational pressure. Larger selective constraints can be observed with increasing N_{e}. This effect is weaker at coevolving sites than at independently evolving sites. The influence of N_{e} on nucleotide variation is not exclusive to tRNAs, but seems to be universal in RNA

molecules as miRNAs exhibit a similar increase of selective constraints with increasing N_e (Table C.S5).

Here, we did not take the effect of recombination on $\overline{T}_{coev}$ into account. It was shown previously that recombination may retard the rate of fixation of compensatory double mutants in RNA molecules even when the distance in sequence (d) between paired nucleotides is small ($50 < d < 250$) (Piskol and Stephan, 2008). For mildly deleterious single mutants, recombination also has the potential to combine individual mutant alleles thus leading to complex adaptations (Lynch, 2010; Weissman et al., 2010). However, usually fixation times of double mutants are only moderately affected by recombination.

While we cannot completely rule out that some of the substitutions investigated in our study are of an adaptive nature, we assumed that the vast majority of mutations in tRNAs are deleterious. Given that tRNA molecules have a well defined function, mutations will most likely alter the structure and original conformation of the molecule in space thus potentially changing its functionality and leading to a decrease in fitness. Very important for our analysis was the assumption that WC base pairs, which form the secondary structure of the tRNA are subject to coevolutionary dynamics, while other nucleotides in the tRNA, whether involved in non-WC pairs or completely unpaired, may evolve independently. This was shown to be the case in bacterial rRNAs (Dutheil et al., 2010) and is also directly applicable to tRNAs due to the universality of base pairs (Leontis and Westhof, 2001).

In a recent study (Piskol and Stephan, 2011) we reported that selective constraints in computationally predicted noncoding RNAs that are encoded in the nuclear genomes of drosophilids and hominids differ in their magnitude between the two genera. We suggested that N_e is responsible for this difference and results in stronger selective constraints in drosophilids. In general, the definition of neutral evolution and the distinction between neutrality and purifying selection in terms of N_e is complicated and has been the topic of many controversies (Nei et al., 2010). Even though the definition of neutrality may have changed over the years (Ohta and Gillespie, 1996; Nei, 2005) our present results demonstrate that N_e indeed can be held accountable for differences in the efficacy of selection and does so by affecting coevolving and independently evolving sites to different degrees. We suggest that there exists a set of peripheral tRNAs for which mutations are slightly deleterious and scaled selective coefficients are only of a moderate size ($|N_e s| \leq 5$). For this regime the pattern of increasing constraints is strongly influenced by the effective population size (Figure 4.3b) and follows the theoretical predictions for fixation times of deleterious mutations in Kimura's one- and two-locus models (Kimura, 1980, 1985). The remaining (core) tRNAs might be subject to stronger evolutionary restrictions and thus

divergence patterns in these molecules are less susceptible to differences in N_e.

Our results have also direct consequences for the inference of phylogenetic relationships between taxa that differ in their long-term effective population size. If the estimation of branch lengths is performed using independently evolving sites that are subject to weak purifying selection (e.g., synonymous sites), then the length of branches leading to taxa with large N_e might be underestimated to a larger extent than for taxa with small N_e.

4.4.1 Acknowledgments

This work was supported by grant Ste 325/8 to W.S. from the Deutsche Forschungsgemeinschaft.

Chapter 5

General Discussion

By combining evolutionary theory, comparative genomics and computational biology this thesis aimed to extend our understanding of the processes that influence the evolutionary dynamics of folded RNA molecules. In Chapter 2 we identified structural and population genetic features of RNA molecules that influence the speed of RNA evolution. Chapter 3 highlighted the importance of ncRNAs based on compelling evidence for strong selective constraints that act upon them. Finally, in Chapter 4, the relation between evolutionary rates of tRNAs and the long-term effective population size of species allowed insights into the difference between evolutionary processes at independently evolving and coevolving loci.

5.1 Structural and Population Genetic Factors

We identified the distance (in nucleotides) between pairing positions, the length of pairing regions, the position of a nucleotide in the pairing region as well as the GC content as determinants of RNA evolution. The secondary structure of an RNA molecule consists of stretches of nucleotides that pair with each other through hydrogen bonds. The result of these pairings is a characteristic folding of the molecule. While each nucleotide by itself is capable of establishing hydrogen bonds with the surrounding solution, the RNA molecule nevertheless favors the folded form. This energetically more stable state does not result

from the sole pairing of the nucleotides, but also from the stacking of adjacent nucleotide pairs (Tinoco et al., 1973). As a result, the integrity of the RNA structure does not only depend on the changes of single base pairs, but also on their impact in the context of neighboring pairs of nucleotides. Therefore, many approaches that were developed to determine the secondary structure of RNAs rely (at least partially) on the energy of tuples of adjacent base pairs for the prediction of the energetically most stable state (Hofacker et al., 1994, 2002; Zuker, 1994). Based on Tinoco's model Mimouni et al. (2009) predicted that the disruption of a nucleotide pair will cause small changes on a local scale but can also lead to major rearrangements in the structure. This effect will differ depending on the position of the base pair in the pairing region. The resulting constraints on the evolution of single pairs are reflected in characteristic nucleotide substitution patterns along the pairing region: penultimate positions in a helix experience the lowest rate of substitution, substitutions occur more often at ultimate positions and are most frequent in inner parts of the helix. Even though our model focused on single nucleotide pairs (without considering their neighborhood), its results concerning the rate of occurrence of covariations, wobbles and mismatches along the helix seem to be in agreement with the findings of Mimouni et al. (2009). Specifically, inner parts of helices are most likely to be populated by wobbles and mismatches, which were able to fix due to the lower constraints against base pair disruptions in that region. These nucleotide variants lead to the high substitution rates in inner parts of helices. The probability for the presence of wobbles and mismatches declines towards helix ends and is lowest at penultimate and ultimate positions. In contrast, the rate of covariation is highest at ultimate positions and declines with larger distance from the helix end. Therefore, compensatory mutations might result in intermediate substitution rates observed at ultimate positions by Mimouni et al. (2009). Their occurrence is probably driven by the following two factors. First, ultimate positions can tolerate base pair disruptions to a certain degree. This results in a shallow valley of reduced fitness that may be crossed by a compensatory mutation. Second, the basic need for base pair formation still persists. It favors the formation of compensatory mutations from intermediates, which ensures the stability of helix ends and thus secures the major building blocks of the RNA secondary structure.

Clearly, our model for the analysis of divergence patterns in ncRNAs still leaves room for improvement. Similar to Tinoco's (1973) model that took tuples of neighboring base

pairs into account and was a great improvement over previous versions that relied on single pairs (Tinoco et al., 1971), also our analysis could benefit from the consideration of the neighboring sequence and the structural neighborhood of base pairs. Furthermore, the exploration of the balance between the occurrence of compensatory mutations, wobbles and mismatches in the context of Mimouni's results would give further insights into the processes that shape nucleotide variation on the intra-helix level.

5.2 Selective Constraints

In Chapters 3 and 4 we calculated selective constraints for nuclear-encoded ncRNAs in drosophilids and hominids and for tRNAs in several vertebrate species. For this purpose we used divergence data from pairs of species (a two lineage approach), and determined the number of substitutions between the sequences without usage of an outgroup. This method is applicable if divergence between species is not large, such that the probability of double hits is low and no fixed mutation events are missed. If the per-site divergence exceeds 0.3, the use of three lineages to polarize substitutions compared to an outgroup species is preferred. For most species pairs under consideration in Chapter 4 the divergence was low enough to safely determine substitutions from two lineages. Only the divergence between *D. melanogaster* and *D. yakuba* as well as chicken and zebra finch was of a magnitude such that multiple hits may have occurred (see Table C.S4). Nevertheless, the two-lineage method was used for these species as well. This allowed us to retain as many tRNAs and neutrally evolving regions as possible on the expense of a slight underestimation of selective constraints. By using three lineages only a smaller number of orthologous tRNAs and intronic positions would have been available for the analysis. This would have made the comparison of the effect of N_e on autosomal and X-linked tRNAs in Drosophila impossible. For the analysis of N_e between species, also the Drosophila pair *D. melanogaster*/*D. simulans* was included. It diverged much more recently and confirmed the overall high constraints in Drosophila without being affected by potential double hits.

Other factors that may influence the estimation of selective constraints are the current GC content of the nucleotide sequence and the GC composition that is expected to be

reached after evolution under constant substitution patterns (equilibrium GC content). It was previously shown that both the GC content and the equilibrium GC content (hereafter noted GC*) vary between species and genomic regions, which results in varying substitution patterns along the genome (Bernardi, 2000; Eyre-Walker and Hurst, 2001; Clément and Arndt, 2011). For instance, deviations from a GC content of 50% lead to increased substitution rates (Előry et al., 2010). The knowledge of GC* allows for the correction of the variation of substitution rates for the calculation of selective constraints (Halligan et al., 2004). However, the correct GC* might not always be known. Nevertheless, the estimation of constraints can be robustly performed if an average GC* is assumed, because constraints are only marginally affected by variation in GC* if selective pressures are high (e.g., constraints at nonsynonymous sites in hominids are in the range of 0.726–0.737 for GC* in the range of 25–50% (Előry et al., 2010)).

The use of an average GC* naturally results in selective constraints that reflect the genomic averages for all ncRNAs (Table 3.3) and tRNAs (Table C.S1) in the investigated species. Apart from the calculation of selective constraints from divergence data the effect of deleterious mutations can be usually characterized by the distribution of fitness effects (DFE). The DFE describes the probability that a new mutation is of adaptive, neutral, slightly deleterious or strongly deleterious nature and is usually calculated from the frequency of nucleotide variants in a population under the assumption of simple demographic scenarios (Eyre-Walker et al., 2006; Eyre-Walker and Keightley, 2007; Keightley and Eyre-Walker, 2007). To date, studies of the DFE have been performed in protein coding and intergenic regions (Eyre-Walker and Keightley, 2009; Gossmann et al., 2010; Keightley and Eyre-Walker, 2010; Slotte et al., 2010; Tellier et al., 2011). The main requisite to calculate the DFE are polymorphism data in neutrally evolving regions (which serve as a reference) and at polymorphic sites under selection for which the DFE is of interest. Unfortunately, polymorphism data for regions that harbor ncRNAs were rare until recently, and publicly available screens for nucleotide variation in humans in form of HapMap (International HapMap Consortium, 2007) and Perlegen (Hinds et al., 2005) data sets have been able to profile only a fraction of polymorphic sites in the human genome. Furthermore, these data sets are subject to ascertainment biases that are often hardly correctable (Nielsen et al., 2004; Clark et al., 2005) and make population genetic analyses that rely on the frequency spectrum of mutations difficult. Due to these restric-

tions the calculation of DFEs for ncRNAs was not feasible until now and limited our approach to the characterization of the deleterious effect of new mutations through comparative genomics of different species. However, recent whole-genome sequencing efforts in the model organisms human (1000 Genomes Project Consortium, 2010), Drosophila and Arabidopsis (Atwell et al., 2010) have yielded population genomic sequence data that are suitable for the estimation of DFEs and will finally allow for the calculation of the distribution of deleterious effects for mutations in ncRNAs. The direct uncovering of the walk of an RNA molecule on a fitness landscape would be even more desirable and was recently demonstrated for an RNA ligase ribozyme (Pitt and Ferré-D'Amaré, 2010). However, the necessary sequencing of millions of RNA molecules still makes the procedure prohibitively expensive to be performed on a large scale and for RNA molecules longer than the typical read length of current high-throughput sequencing techniques (∼75 nucleotides). Therefore, also in the near future the calculation of selective constraints from divergence data and the estimation of DFEs will remain the method of choice for the identification of the depth of valleys in the fitness landscapes of RNA molecules.

5.3 The Role of the Effective Population Size

We were able to show in Chapter 4 that the speed of evolution in tRNAs decreases with an increase of the effective population size (N_e) and that independently evolving sites are affected to a larger extent by differences in N_e than coevolving positions. To investigate the effect of different N_e we assumed that the selection coefficient (s) is constant between species and that the efficacy of selection only differs due to differences in the effective population size. However, it is unlikely that s, which is usually quite small in nuclear encoded RNAs (on the order of $10^{-4}-10^{-5}$), would remain constant over long evolutionary periods and would be the same for different pairs of species. Nevertheless, the patterns described in Chapter 4 suggest that the speed of evolution in our data depends on the systematic increase in N_e.

In our study, we used previously determined long-term effective population sizes for all vertebrate and Drosophila species pairs. In many of these cases, the long-term N_e was calculated as an average of the individual effective population sizes, which were obtained

from the nucleotide diversity of each of the two species. These approximations and the fact that diversity in the domesticated species in our data (dog, cat, and chicken) differs from that in their wild relatives, prohibited a quantitative use of the estimates of N_e. Nevertheless, N_e differs by orders of magnitude between hominid, murid, and drosophilid species pairs, which allowed for a semi-quantitative analysis of its effect.

Our observations of the relation between selective constraints and N_e are based on the idea that slightly deleterious mutations are observed if N_e is large, but may be regarded as neutral in species of small N_e. In Chapter 4, we portray our results in the context of Kimura's work. The primary purpose of our study was not to determine the definition of neutrality in RNA molecules according to N_e but simply the identification of the role of N_e in their evolution. Especially, we were interested in the difference of its impact on independently evolving and coevolving sites. Regardless of the definition of neutrality, we were able to determine that N_e influences the efficacy of selection and could identify two sets of tRNAs that seem to be subject to different levels of selection. One set shows a large impact of N_e and is probably only experiencing weak purifying selection, while the second set of tRNAs barely shows signs of the impact of N_e, which agrees with theoretical predictions in the case of strong purifying selection.

5.4 Future Directions

Apart from the work described above, several interesting aspects of RNA evolution also deserve future attention. For instance, we mainly investigated the idea that pairing regions are subject to evolutionary constraints because nucleotide pairs therein should remain intact. However, the pressure on each nucleotide to remain unchanged is also exerted to avoid putative new pairings that might be formed between the mutated position and other sites than the original partner. Similarly, unpaired regions are probably under an evolutionary pressure that prevents changes, which would enable alternative pairings. In this respect it would be of interest to determine whether the molecule has reached a region of the evolutionary landscape where small changes in the sequence will result in large changes in the conformation (and thus phenotype) of the molecule, or whether it is rather in a state where nucleotide changes rarely lead to structural alterations. Relations

between robustness (the frequency of mutations that do not alter the phenotype) and evolvability (the speed at which a phenotype moves along the fitness landscape) have been studied for RNAs mostly from a theoretical perspective (Ancel and Fontana, 2000; Cowperthwaite et al., 2008; Parter et al., 2008; Elena and Sanjuán, 2008). Their interplay is complex and the evolutionary outcome not always intuitive. For instance, population genetic models suggest that the speed at which RNA landscapes are being explored relates to the robustness of the molecule in a non-monotonic fashion, but also depends on the fraction of equally likely phenotypes that it can access (Draghi et al., 2010). These models demand an experimental and empirical examination to determine the connections between robustness and evolvability and could reveal interesting insights into the routes that an RNA molecule might have taken on the evolutionary landscape.

Furthermore, our work focused on intramolecular interactions in RNAs and their influence on patterns of divergence, which reflect the selective constraints the molecule is subject to. However, intramolecular factors are not the only ones that have an effect on the rate of substitution. RNA molecules are almost always involved in non-local interactions with proteins or other RNAs. These interactions can be sequence- and structure-specific. A good example is the structure-specific recognition of the pre-miRNA during its maturation process and the sequence-specific recognition of target sites by the miRNA itself. Both have an influence on the rates of evolution in different parts of the pre-miRNA and lead to specific substitution patterns along the RNA molecule (Ma et al., 2010). Other non-local interactions include the localization signals in 3'UTRs of the *bicoid* gene in Drosophila, long-range interactions in the *Adh* mRNA (Baines et al., 2004) or interactions in tRNAs where parts of the structure act as recognition sites for the ribosome (T-loop) or for aminoacyl synthetases that link an amino acid to the tRNA (acceptor stem). In such complex interactions mutations may result in a complete absence of function (e.g., failure of correct localization of the *bicoid* mRNA (Irion and Johnston, 2007)) or lead to gradually changing phenotypes. A good example for the latter case is the post-transcriptional modification of RNAs (RNA editing), which results in varying levels of transcripts of the same type.

5.5 RNA editing

RNA editing is the process in which the original information of an RNA molecule is altered by post-transcriptional enzymatic modification. In humans, the two known types of editing that lead to changes in nucleotides are Adenosine-to-Inosine (A-to-I) and Cytosine-to-Uracil (C-to-U), catalyzed by adenosine deaminases acting on RNA (ADARs) and apolipoprotein B mRNA editing enzyme APOBEC1, respectively (Bass, 2002; Blanc and Davidson, 2003). Also other types of editing that lead to differences between genomic DNA and transcribed RNA are believed to exist, but currently lack experimental evidence. Most editing events occur in Alu repeats. Nevertheless, increasing numbers of transcripts are reported to be subject to RNA recoding, which leads to amino acid substitutions in the resulting protein (Yamanaka et al., 1997). RNA editing has the potential to largely increase genome complexity and fine-tune RNA functions (Grohmann et al., 2010). Extreme examples are the PIGO and RAB27A genes, which were shown to contain 42 and 50 editing sites that result in 832 and 401 different transcripts, respectively (Paz-Yaacov et al., 2010). Levels of editing can vary from a few edited copies to editing of all transcripts (Li et al., 2009). However, the factors that regulate the exact level of editing are still not known. Nevertheless, changes of editing levels in mice, which were shown to result in serious neurological diseases (Brusa et al., 1995) and in the lethality of ADAR1/2 deficient mice (Riedmann et al., 2008), emphasize the functional importance of the editing process. Furthermore, the increased level of RNA editing in human brain tissues compared to other primates and its association with neuronal functions (Hoopengardner et al., 2003) hint at the possible contribution of A-to-I editing to the development of higher brain functions in modern humans (Paz-Yaacov et al., 2010).

One of the best studied examples of RNA editing by APOBEC1 (even though by far not the only one) is the C-to-U modification of the spliced and polyadenylated apolipoprotein B (apoB) mRNA. In this process, a CAA codon is changed into a UAA stop codon, which results in the production of a shortened protein with altered function (Anant and Davidson, 2001). The efficiency of recruitment and the assembly of the complex that performs the modification depends mainly on two factors (Figure 5.1): 1) a primary sequence motif downstream of the edited site (Shah et al., 1991), and 2) the folding of the region

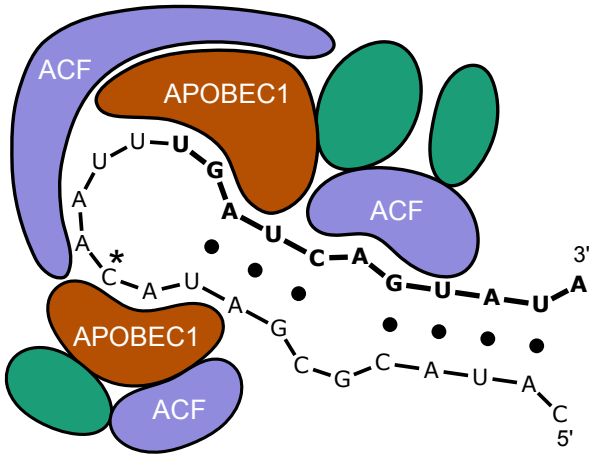

FIGURE 5.1. **C-to-U RNA editing of apolipoprotein B.** APOBEC1 and its complementation factor (ACF) bind to a region surrounding the base to be edited (asterisk) at position 6666 of the apolipoprotein B mRNA. The formation of a hairpin structure and the presence of a specific sequence motif (bold) are necessary for the recruitment and correct function of the editing complex. (Figure adapted from Blanc and Davidson (2003))

into a distinct hairpin that exposes the edited site to the enzyme complex (Hersberger et al., 1999). Therefore, the sequence that encompasses the edited site has to fulfill two concurrent functions, as it is involved in intramolecular interactions but also in interactions with trans-acting factors. This marks an interesting system, in which the balance between pressures on structure and sequence conservation may be investigated. While it is known that changes in either one of the factors alter the efficiency of editing and thereby the number of edited transcripts, the regulation of the exact level is still largely unknown. The comparisons of C-to-U editing levels at different genomic positions and between species may help to determine the structural and sequence related factors that constitute the exact level of editing and may also reveal the importance of long-range interactions that are necessary for the process.

5.6 Conclusions

The process of RNA recoding is only one of many examples that demonstrate the complexity of the interactions RNA molecules are involved in. In this sense, a study of structural and population genetic factors of RNA molecules is only able to uncover a subset of all determinants of RNA sequence and structure evolution. Nevertheless, the factors that

were studied here describe the most basic forces that leave their footprint in the nucleotide variation that can be observed in RNA molecules. These forces are universal between classes of RNAs and also extend across genera. Even though the processes that are responsible for shaping nucleotide divergence in RNA molecules were studied using between-species comparisons, all of them have their roots on a population genetic level. From there, nucleotide variants are spread to a species wide level. Their fixation process is mainly determined by the factors described here, but also by unique functional properties and therefore also by constraints that are unique to each class of RNAs. Our study was able to shed light onto the factors common to all RNAs, while the specific factors have to be identified for each RNA class separately.

Relative change in odds

$$\begin{aligned}
\frac{\gamma(x+1)}{\gamma(x)} &= \frac{P(Y=1|x+1)}{P(Y=0|x+1)} \cdot \frac{P(Y=0|x)}{P(Y=1|x)} \\
&= \frac{\pi(x+1)}{1-\pi(x+1)} \cdot \frac{1-\pi(x)}{\pi(x)} \\
&= \frac{exp(\beta_0 + (x+1)\beta)}{exp(\beta_0 + x\beta)} \\
&= \frac{exp(\beta_0) \cdot exp(x\beta) \cdot exp(\beta)}{exp(\beta_0) \cdot exp(x\beta)} \\
&= exp(\beta)
\end{aligned}$$

Correlation between influence variables in logistic regression

	x_1	x_2	x_3	x_4	x_5
x_1	–	0.0606 (8.679×10^{-7}***)	0.0426 (0.0055***)	0.0482 (9.43×10^{-5}***)	−0.0299 (0.0158*)
x_2	–	–	0.5344 ($< 2.2 \times 10^{-16}$***)	0.0728 (3.391×10^{-9}***)	−0.0549 (8.38×10^{-6}***)
x_3	–	–	–	0.0458 (0.0002***)	−0.0353 (0.0042***)
x_4	–	–	–	–	0.1035 ($< 2.2 \times 10^{-16}$***)
x_5	–	–	–	–	–

Table A.S1. **Correlation between variables in logistic regression.** Here, columns that contain covariations or lack any substitution event from *DS1* were used (6571 entries). Variables $x_1, ..., x_5$ represent the distance, helix length, average distance to helix end, average substitution rate and GC content, respectively.

Appendix A.

Substitution rates in helices of non-miRNA and miRNA folds

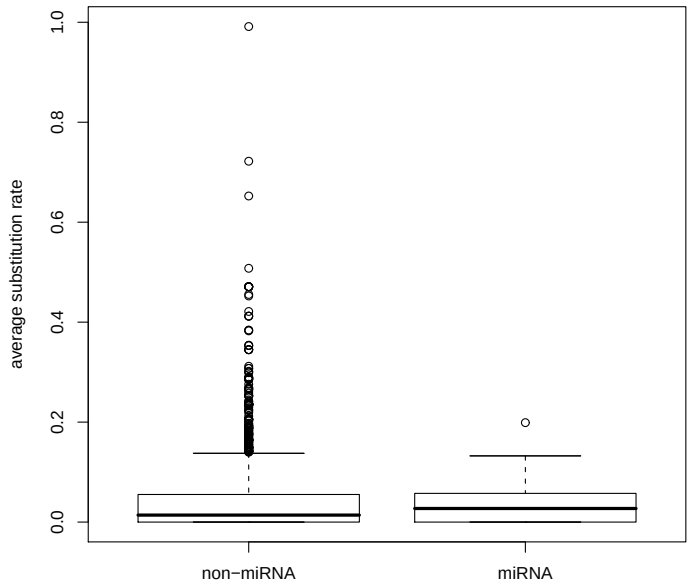

Figure A.S1. **Average substitution rates in helices in non-miRNA folds and in folds that were found to overlap with miRNA predictions based on the UCSC Genome Browser annotation**

	non-miRNA	miRNA
Q1-1.5*IQR	0.00000	0.00000
Q1	0.00000	0.00000
median	0.01396	0.02716
Q3	0.05520	0.05744
Q3+1.5*IQR	0.13759	0.13246
mean	0.04135	0.04341
observations	3064	52

Table A.S2. **Summary statistics of the average substitution rate for helices in non-miRNA folds and helices in folds annotated as miRNAs.** Q1 and Q3 represent the first and third quartile. IQR is the interquartile range of the distributions of average substitution rates for both classes. 489 folds (3064 helices) were classified as non-miRNA, while 18 folds (52 helices) overlap with miRNA predictions.

VIF values for all models

	DS1	*DS2*
Distance	1.004824	1.016749
Helix length	1.238444	1.142450
Distance to helix end	1.213730	1.124105
Avg. subst. rate	–	–
GC content	1.031231	1.011479

Table A.S3. **VIF values for the estimated coefficients in *DS1* and *DS2* for covariations (from Table 1 main text).**

	DS1 (wobbles)	*DS1* (mismatches)
Distance	1.003933	–
Helix length	1.438030	1.514042
Distance to helix end	1.437295	1.484634
Avg. subst. Rate	1.002448	1.010699
GC content	–	1.024458

Table A.S4. **VIF values for the coefficients in *DS1* for wobbles and mismatches (from Table 3 main text).**

Appendix A.

Substitution Rates in Paired and Unpaired Regions

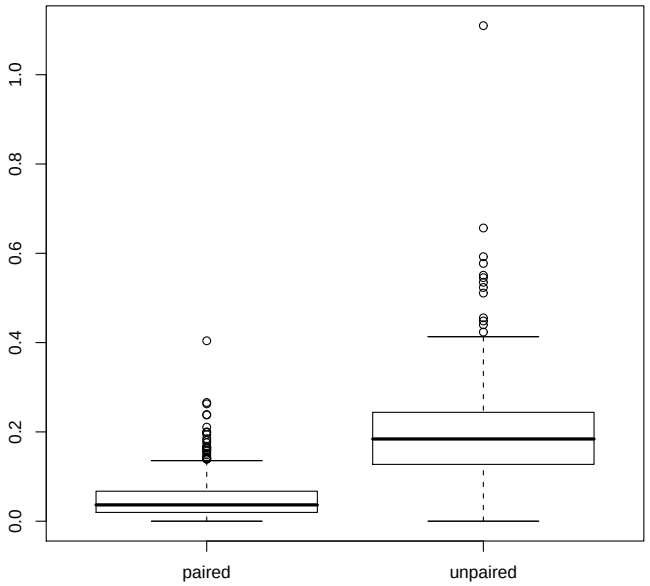

Figure A.S2. **Boxplot of average substitution rates in paired and unpaired regions of all 507 folds in *DS1***

	paired	unpaired
Q1-1.5*IQR	0.00000	0.00000
Q1	0.01969	0.12725
median	0.036554	0.18418
Q3	0.067197	0.24385
Q3+1.5*IQR	0.135652	0.41326
mean	0.05036	0.19694

Table A.S5. **Summary statistics for the average substitution rates in paired and unpaired parts of the folds in *DS1*.** The difference between the distributions of substitution rates is highly significant (Pearson: $\rho = 0.2881673$, $P = 3.757 \times 10^{-11}$)

Testing the Influence of RNA Structure Prediction Algorithms on the Observed Patterns of Nucleotide Variation

To identify biases in secondary structure prediction by EvoFold (Pedersen et al., 2006) that may be held accountable for the putative impact of structure prediction on the observed patterns of compensatory mutations, we simulated sequences with known secondary structure and compared nucleotide variation based on the known structure to variation based on a structure predicted by EvoFold. Furthermore, we tested whether the alignment method is a potential source for the observed pattern by re-aligning sequences from *DS1* and re-annotating the structures using various methods for alignment and secondary structure prediction. In addition shuffled alignments were obtained to ensure that previous findings are not just an artifact of a random process.

To rule out that the substitution pattern under investigation is introduced by a bias in secondary structure prediction by EvoFold, we randomly simulated 5000 secondary structures and sequences via stochastic context free grammars (SCFGs) (Eddy and Durbin, 1994) using a program written by Dirk Metzler. The resulting sequences were of an average length of 350 nucleotides and dinucleotide contents of GC: 25%, AU: 65%, GU: 10%. Each sequence was subjected to an evolutionary process along the phylogenetic tree corresponding to the 8 species in the original data (fig. A.S3a). For this purpose a C++ program written by Dirk Metzler was ported to JAVA and extended to simulate the evolution of dinucleotide pairs. For the evolution of unpaired regions we used the F84 model (Felsenstein, 1984), while evolution for paired regions was performed based on substitution rates derived from the pfold rate matrix (Knudsen and Hein, 1999, 2003). The resulting alignment served as input for secondary structure prediction by EvoFold after discarding the original information about the folded structure (fig. A.S3b). Subsequently, two logistic regression models for covariations as response variables were calculated. One was based on simulated structures the other on predicted EvoFold structures (Table A.S6). Since base pairs were simulated independently of each other, the simulated structures do not show any significant effects in the estimates of distance between pairing nucleotides, helix length and distance to helix end. They serve as a reference when determining a bias in structure prediction by EvoFold, which then should be visible in the significant estimates for the predicted structures. Indeed, the EvoFold predictions show significant estimates for the distance between pairing nucleotides, helix length, distance to the helix end and GC content. However, compared to the estimates obtained from the real data

(Table 1 main text), the signs for the former three estimates are reversed. Hence, we can conclude that the previously found estimates for these three parameters are conservative (e.g., the originally determined negative estimate of −0.1914 for the distance to the end of a helix should be even lower, given the positive bias in EvoFold predictions).

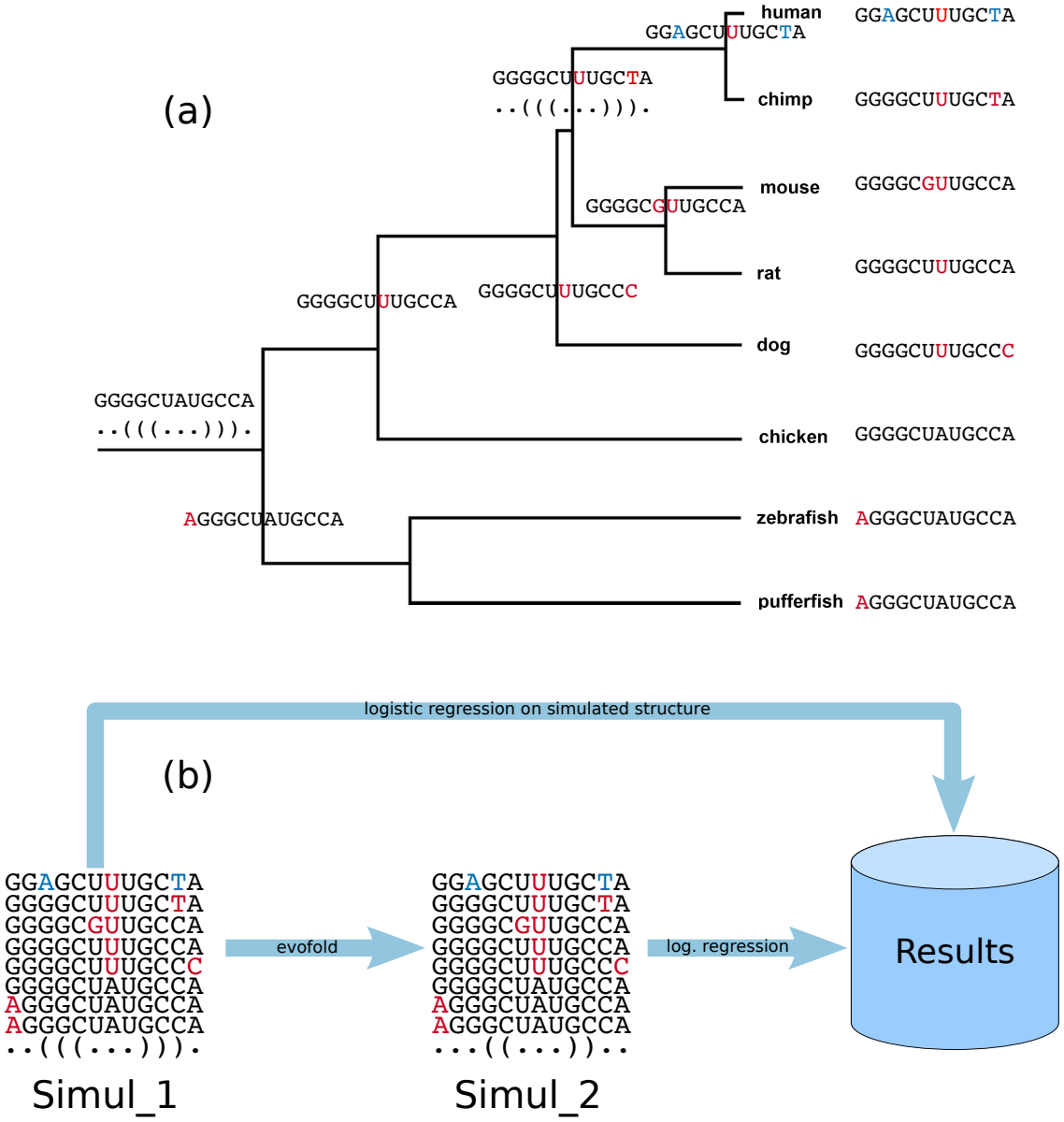

FIGURE A.S3. **Simulation procedure to determine potential biases of secondary structure prediction on estimates obtained from the logistic regression.** 1) Secondary structures and associated sequences were randomly generated (n=5000, avg. length=350nt, GC:25%, AT:65%, GT:10%). 2) Each sequence was evolved along the phylogenetic tree using appropriate models of sequence evolution for paired and unpaired regions (fig. A.S3a). 3) Logistic regression analysis was performed on the resulting alignments taking the simulated structures (*Simul_1*) and structures predicted by EvoFold (*Simul_2*) into account (fig. A.S3b).

	Simul_1		*Simul_2*	
Independent variable	Estimate	Pr(> |z|)	Estimate	Pr(> |z|)
Intercept	−0.5924	$< 2 \times 10^{-16}$***	−0.7453	$< 2 \times 10^{-16}$***
Distance	0.0000	0.3600	0.0005	2.72×10^{-11}***
Helix length	−0.0004	0.7270	−0.0124	$< 2 \times 10^{-16}$***
Distance to helix end	0.0047	0.1320	0.0348	$< 2 \times 10^{-16}$***
Avg. subst. rate	−0.0801	0.2750	−0.1386	0.2530
GC content	0.2097	3.09×10^{-14}***	0.5727	$< 2 \times 10^{-16}$***

TABLE A.S6. **Minimal logistic regression models for simulated and predicted structures**

Subsequently, we discarded secondary structure and alignment information from all 507 folds in *DS1* and applied various combinations of alignment and secondary structure prediction, to ensure that our findings are not specific for the methodology used to acquire the structures. For alignment purposes we used progressive: ClustalW (Thompson et al., 1994); iterative: muscle (Edgar, 2004), mafft (Katoh et al., 2002), stemloc (Holmes, 2005) and consistency based: DIALIGN-TX (Subramanian et al., 2008), t_coffee (Notredame et al., 2000), m_coffee (Wallace et al., 2006) approaches. In addition a shuffled version of the original alignment was created in which no significant pattern should be visible. Subsequently, secondary structures were predicted on these alignments by Pfold (Knudsen and Hein, 1999, 2003). Since EvoFold relies on the same methodology as Pfold which is based on SCFGs we also included the thermodynamic method RNAalifold (Hofacker et al., 2002) as well as PETfold (Seemann et al., 2008), which combines phylogenetic and thermodynamic information. We also used the combination of RNAshapes (Steffen et al., 2006) and RNAforester (Höchsmann et al., 2004), which aligns previously determined common shapes as well as mlocarna (Will et al., 2007) for simultaneous alignment and structure determination. Logistic regression models were then calculated based on the resulting alignments and structures and their estimates compared to the values that were obtained by the EvoFold algorithm. For all three variables of interest (distance, helix length, and distance to helix end) most combinations of alignment and secondary structure predictors yield estimates and significance values that match the originally obtained ones (light gray bars in fig. A.S4). This is true for structures predicted with Pfold and PETfold (which are based on algorithms related to EvoFold), but also RNAalifold structures result in estimates of comparable magnitude and sign even though not significant in all the cases (dark gray bars in fig. A.S4). The shuffled alignments lead either to estimates that differ in sign, or are statistically not significant in most cases (black bars in fig. A.S4).

Therefore, we can conclude that the observed patterns are free from biases in secondary structure prediction by EvoFold and independent of the methodology of alignment and structure prediction. These findings also indicate that the decreasing rate of covariations with growing distance between paired nucleotides and growing distance from the helix end as well as an increased rate of covariations in longer helices are consequence of an evolutionary process and not randomly generated.

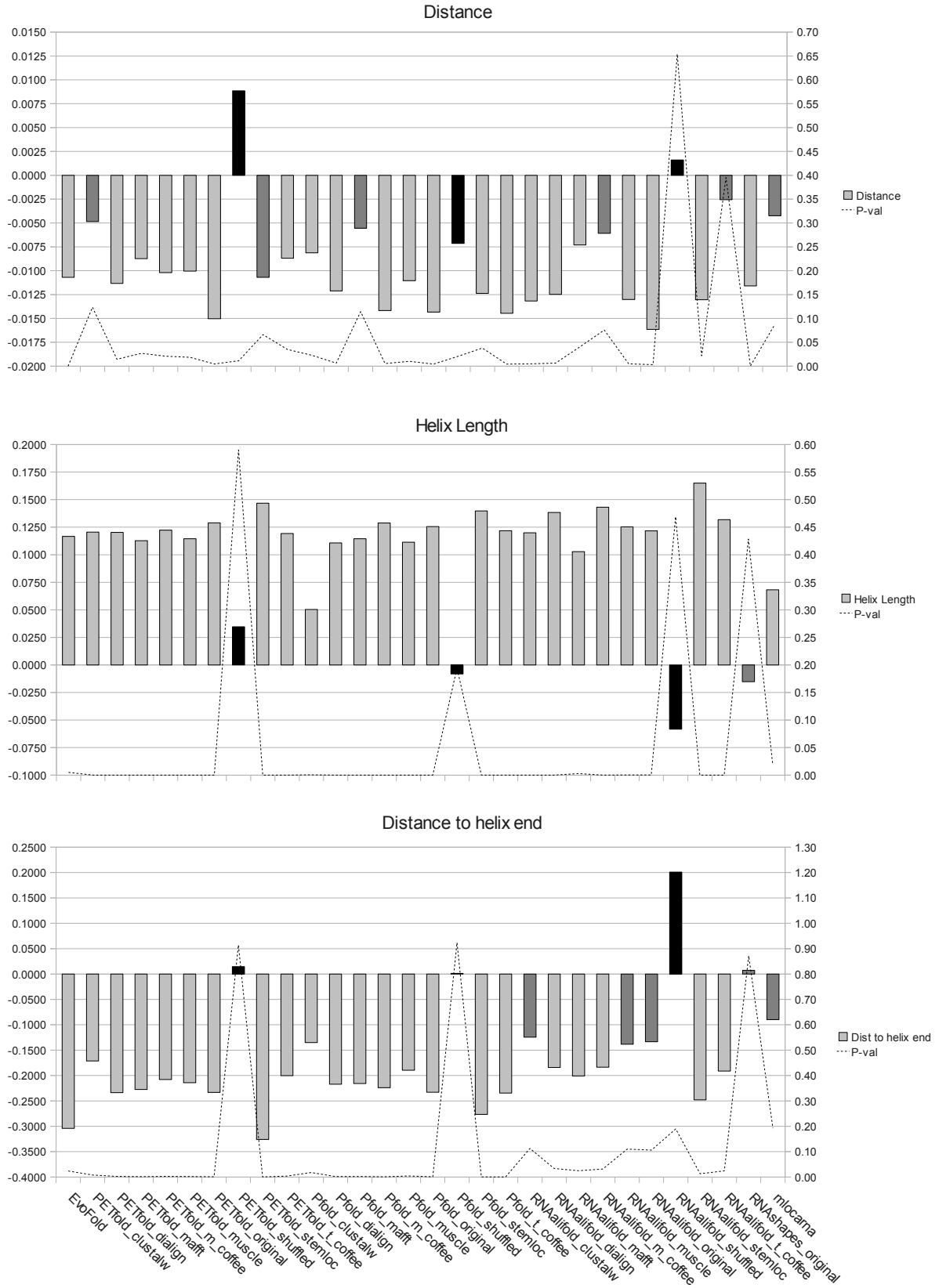

FIGURE A.S4. **Logistic regression estimates for 1) distance, 2) helix length, and 3) distance to helix end, based on structures and alignments that were obtained using different methodologies.** The leftmost bar (labeled EvoFold) represents the estimates for *DS2* from Table 1 (main text). The remaining light gray bars show estimates that exhibit the same direction of influence as the original data and are significant on the 0.05 level. Not significant estimates are shown in intermediate gray color. Black bars represent estimates obtained from shuffled alignments.

Substitution Rates

	Sequence Type	k_{paired}	(95% CI)	k_{unpaired}	(95% CI)
A.	Drosophilids				
	intron_UTR_ST	0.0010	(0.0002, 0.0028)	0.0131	(0.0101, 0.0166)
	intron_UTR_AS	0.0011	(0.0004, 0.0020)	0.0116	(0.0089, 0.0143)
	intron_nUTR_ST	0.0008	(0.0003, 0.0014)	0.0112	(0.0094, 0.0132)
	intron_nUTR_AS	0.0014	(0.0007, 0.0021)	0.0128	(0.0110, 0.0149)
	exon_nUTR_ST	0.0067	(0.0041, 0.0097)	0.0227	(0.0185, 0.0272)
	exon_nUTR_AS	0.0043	(0.0022, 0.0067)	0.0137	(0.0112, 0.0167)
	exon_UTR_ST	0.0018	(0.0004, 0.0036)	0.0164	(0.0118, 0.0215)
	exon_UTR_AS	0.0025	(0.0011, 0.0044)	0.0172	(0.0130, 0.0212)
	intergenic	0.0014	(0.0011, 0.0017)	0.0122	(0.0116, 0.0129)
	intronic ≤65nt (8-30)	-	-	0.1100	(0.1074, 0.1125)
B.	Hominids				
	intron_UTR_ST	0.0010	(0.0000, 0.0031)	0.0066	(0.0019, 0.0116)
	intron_UTR_AS	0.0000	(0.0000, 0.0000)	0.0063	(0.0031, 0.0102)
	intron_nUTR_ST	0.0006	(0.0002, 0.0013)	0.0032	(0.0019, 0.0044)
	intron_nUTR_AS	0.0005	(0.0001, 0.0013)	0.0031	(0.0018, 0.0046)
	exon_nUTR_ST	0.0008	(0.0005, 0.0013)	0.0039	(0.0029, 0.0048)
	exon_nUTR_AS	0.0007	(0.0003, 0.0013)	0.0032	(0.0023, 0.0043)
	exon_UTR_ST	0.0003	(0.0000, 0.0008)	0.0027	(0.0014, 0.0041)
	exon_UTR_AS	0.0007	(0.0002, 0.0014)	0.0030	(0.0016, 0.0046)
	intergenic	0.0007	(0.0004, 0.0010)	0.0045	(0.0016, 0.0046)
	intronic AR	-	-	0.0104	(0.0102, 0.0105)
	intergenic AR	-	-	0.0118	(0.0116, 0.0119)

Table B.S1. **Substitution rates and CIs for paired and unpaired sites in sequences of different types.** Note.– Rates for folded regions are given at paired (k_{paired}) and unpaired positions (k_{unpaired}). Substitution rates for sequences used as neutral standards are given in the k_{unpaired} column.

Appendix B.

Likelihood Ratio Test for Model Choice

We compared the log likelihoods of three models of nucleotide evolution (K80, HKY85, REV) (Table B.S2) using the likelihood ratio test to determine the model that fits the data best (taking the difference in the number of degrees of freedom (df) that have to be estimated in each model into account). Models HKY85 and REV perform significantly better than model K80 in all cases, given differences in df = 3 and 7, respectively (Table B.S3). Overall the REV model performs better than HKY85 (diff. df = 4) and was chosen for the main analysis.

	Sequence Type	paired			unpaired		
		K80	HKY85	REV	K80	HKY85	REV
A.	Drosophilids						
	intron_UTR_ST	-7449.9777	-6810.6007	-6805.4008	-12503.2114	-12488.2028	-12484.8809
	intron_UTR_AS	-9839.4541	-9071.4065	-9068.0499	-15949.5984	-15926.7870	-15923.3717
	intron_nUTR_ST	-16386.4532	-15109.0507	-15102.4796	-25106.2767	-25061.3396	-25057.0697
	intron_nUTR_AS	-18630.5080	-17175.6329	-17170.3526	-30702.0036	-30768.9410	-30659.1432
	exon_nUTR_ST	-7766.2337	-7740.6018	-7731.1806	-12361.5298	-12434.4603	-12268.2911
	exon_nUTR_AS	-11667.3664	-11566.3928	-11562.2854	-17938.6681	-17849.3713	-17822.1233
	exon_UTR_ST	-4404.7441	-4157.8128	-4154.2369	-7554.2467	-7520.0363	-7508.7868
	exon_UTR_AS	-7461.7313	-7222.7137	-7221.6118	-10968.0496	-10927.5637	-10912.8517
	intergenic	-167815.5256	-154681.1396	-154613.3655	-293888.7201	-288817.1601	-288797.8151
	intronic $\leq$ 65nt (8-30)	-	-	-	-376938.1557	-367959.5888	-367762.3275
B.	Hominids						
	intron_UTR_ST	-1964.4661	-1887.2480	-1883.7767	-4351.4568	-4335.1017	-4330.2107
	intron_UTR_AS	-3922.7438	-3755.6171	-3749.4157	-7786.9958	-7756.3825	-7750.6703
	intron_nUTR_ST	-18109.8433	-17439.2167	-17410.7673	-33757.8555	-33657.4649	-33636.3613
	intron_nUTR_AS	-12956.5403	-12503.3040	-12481.3610	-23797.2389	-23752.7713	-23708.9566
	exon_nUTR_ST	-54460.7114	-54152.0780	-54013.9953	-77351.3199	-76462.2328	-76250.3336
	exon_nUTR_AS	-31049.7961	-30896.3383	-30818.2690	-46196.1803	-45820.3331	-45700.5035
	exon_UTR_ST	-14375.6256	-14010.7442	-14002.6043	-23978.2853	-23657.2455	-23642.8513
	exon_UTR_AS	-13983.3977	-13749.5828	-13747.1973	-23602.9995	-23415.1720	-23404.5554
	intergenic	-76762.6618	-71608.8948	-71491.7925	-142673.3306	-142364.7741	-142286.6263

Table B.S2. **Log likelihood (lnL) values for model estimates at paired and unpaired sites of drosophilid and hominid folds and at sites used as neutral standards in drosophilids.** Note.– lnL values for neutral sites in drosophilids (intron $\leq$65nt (8-30)) are given in the columns for unpaired nucleotides. lnL values for neutral sites in hominids (intergenic ARs/intronic ARs) are not shown. Substitution rates for these sites were calculated in windows of 1-Mb along the chromosomes. Hence, lnL values exist for each of the windows.

Sequence Type	paired			unpaired		
	K80 vs. HKY85	K80 vs. REV	HKY85 vs. REV	K80 vs. HKY85	K80 vs. REV	HKY85 vs. REV
A.Drosophilids						
intron_UTR_ST	5.995×10^{-277}	3.691×10^{-274}	0.034	1.369×10^{-6}	5.438×10^{-06}	0.156
intron_UTR_AS	0	0	0.152	0	4.752×10^{-09}	0.145
intron_nUTR_ST	0	0	0.011	2.331×10^{-19}	2.292×10^{-18}	0.074
intron_nUTR_AS	0	0	0.032	1	9.332×10^{-16}	2.291×10^{-46}
exon_nUTR_ST	4.298×10^{-11}	1.406×10^{-12}	8.440×10^{-4}	1	8.338×10^{-37}	1.140×10^{-70}
exon_nUTR_AS	1.601×10^{-43}	8.063×10^{-42}	0.084	1.775×10^{-38}	1.095×10^{-46}	4.143×10^{-11}
exon_UTR_ST	1.020×10^{-106}	4.851×10^{-104}	0.128	9.297×10^{-15}	8.008×10^{-17}	1.594×10^{-4}
exon_UTR_AS	2.745×10^{-103}	1.417×10^{-99}	0.698	1.899×10^{-17}	7.600×10^{-21}	6.410×10^{-6}
intergenic	0	0	2.532×10^{-28}	0	0	8.073×10^{-08}
intronic <= 65nt (8-30)	-	-	-	0	0	4.244×10^{-84}
B.Hominids						
intron_UTR_ST	2.909×10^{-33}	1.644×10^{-31}	0.139	3.707×10^{-7}	4.179×10^{-7}	0.044
intron_UTR_AS	3.289×10^{-72}	6.403×10^{-71}	0.0146	3.214×10^{-13}	4.296×10^{-13}	0.022
intron_nUTR_ST	1.647×10^{-290}	9.693×10^{-298}	1.299×10^{-11}	2.860×10^{-43}	8.600×10^{-49}	1.511×10^{-8}
intron_nUTR_AS	3.492×10^{-72}	6.384×10^{-201}	6.775×10^{-9}	3.215×10^{-13}	1.035×10^{-34}	4.197×10^{-18}
exon_nUTR_ST	1.820×10^{-133}	1.258×10^{-188}	1.495×10^{-58}	0	0	2.002×10^{-90}
exon_nUTR_AS	3.169×10^{-66}	6.977×10^{-96}	9.839×10^{-33}	1.295×10^{-162}	8.890×10^{-210}	1.099×10^{-50}
exon_UTR_ST	7.381×10^{-158}	8.120×10^{-157}	0.003	7.596×10^{-139}	1.314×10^{-140}	8.630×10^{-6}
exon_UTR_AS	4.935×10^{-101}	6.850×10^{-98}	0.312	4.150×10^{-81}	1.109×10^{-81}	2.847×10^{-4}
intergenic	0	0	1.642×10^{-49}	1.965×10^{-133}	1.014×10^{-162}	9.105×10^{-33}
intronic AR	-	-	-	0.027	0.032	0.775
intergenic AR	-	-	-	3.467×10^{-4}	6.935×10^{-4}	0.127

TABLE B.S3. **LRT *P*-values.** NOTE.– The difference in degrees of freedom was taken as 3, 7 and 4 for the comparisons of K80/HKY85, K80/REV and HKY/REV, respectively. Values for intronic ARs and intergenic ARs represent the fraction of 1-Mb windows for which the LRT test was not significant on the 5% level (e.g., for 3.2% of all windows containing intronic ARs the REV model did not give significantly better lnL values than the K80 model).

Appendix B.

chromosome	codon position		
	1	2	3
chr1	0.390	0.219	0.085
chr2	0.306	0.280	0.376
chr3	-0.303	-0.045	0.250
chr4	1.000	1.000	1.000
chr5	0.528	0.682	0.209
chr6	0.521	0.416	-0.146
chr7	0.742	0.747	0.235
chr8	0.573	0.575	0.165
chr9	0.424	-0.149	0.281
chr10	1.000	1.000	1.000
chr11	0.143	0.161	-0.048
chr12	-0.158	0.509	-0.136
chr13	1.000	1.000	1.000
chr14	0.508	0.455	0.385
chr15	1.000	1.000	0.412
chr16	0.357	0.533	0.491
chr17	0.002	0.390	-0.167
chr18	0.440	0.379	0.111
chr19	0.767	0.314	0.275
chr20	1.000	1.000	0.374
chr21	0.698	0.718	-0.111
chr22	1.000	0.507	1.000
chrX	0.600	0.749	0.694
chrY	1.000	-2.613	-0.491
average	0.564	0.409	0.302

TABLE B.S4. **Constraint (C) at first, second, and third codon positions of single transcript protein coding genes with exactly one exon.** NOTE.– Constraints were calculated from the comparison of human and chimpanzee CDS in genomic regions that were subject to the same selection criteria as the RNA molecules in our study (see Materials and Methods section 3.2.3). The value at the first codon position is slightly lower than at 0-fold sites ($C_{\text{0-fold/ST}} = 0.698$ (Előry et al., 2010)) since this position does not contain 0-fold sites exclusively. Accordingly, constraint at the third codon position is slightly higher than at 4-fold sites ($C_{\text{4-fold/ST}} = 0.215$ (Előry et al., 2010)) since third codon positions may also contain other than 4-fold sites. The chromosome annotation is given according to the location in the human reference sequence.

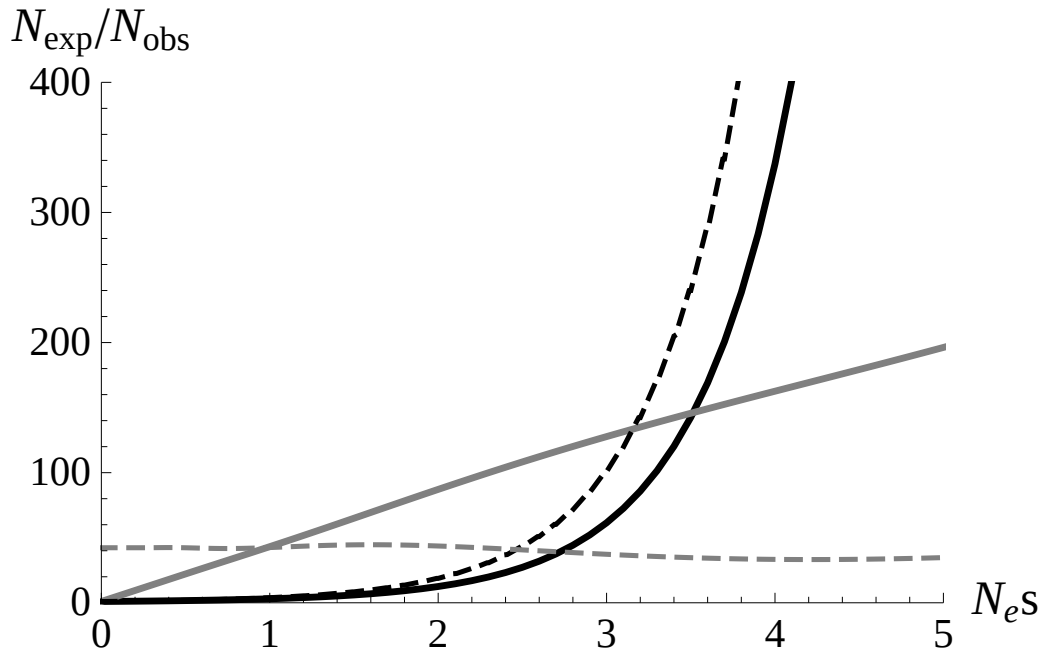

FIGURE B.S1. **Ratio of $N_{\mathbf{exp}}$ to $N_{\mathbf{obs}}$ as a function of $N_{\mathrm{e}}s$ for independently evolving sites (black solid line) and coevolving sites (gray solid line)**. The dashed black and gray lines show the corresponding slopes. For $N_{\mathrm{e}}s \in [0.0, 2.5]$ the ratio increases faster at coevolving sites (larger slope) than at independently evolving sites and thus results in smaller differences between constraints ($C = 1 - \frac{N_{\mathrm{obs}}}{N_{\mathrm{exp}}}$) at paired sites than unpaired sites in the comparison of drosophilids and hominids.

NOTE.–The ratio for coevolving sites (gray line) was obtained from eqn. 16 of Kimura (1985). The ratio for independently evolving sites (black line) was obtained from Kimura's (1980) eqn. 13. In both cases $N_{\mathrm{e}}s = 100,000$ and $\theta = 0.03$ were used to approximately match the parameters assumed in *D. melanogaster*.

	Sequence Type	drosophilids					hominids		
				$N_e s$					$N_e s$
		$\frac{N_{exp}}{N_{obs}}$	95% CI	($\theta = 0.03$)	($\theta = 0.01$)	($\theta = 0.003$)	$\frac{N_{exp}}{N_{obs}}$	95% CI	($\theta = 0.001$)
A.	paired vs. neutral								
	intron_UTR_ST	93.246	(42.484, 469.867)	3.368	2.083	1.875	9.516	(3.142, NaN)	**1.088**
	intron_UTR_AS	89.079	(51.271, 307.243)	3.251	2.057	1.860	NaN	(NaN, NaN)	NaN
	intron_nUTR_ST	114.900	(68.866, 340.586)	**3.994**	**2.214**	**1.949**	14.287	(7.189, 42.931)	1.220
	intron_nUTR_AS	73.094	(48.326, 130.850)	2.825	1.952	1.792	15.977	(7.148, 64.113)	1.255
	exon_nUTR_ST	15.601	(11.271, 24.774)	**1.504**	**1.327**	**1.288**	12.229	(8.423, 21.992)	1.169
	exon_nUTR_AS	38.113	(25.875, 69.321)	2.038	1.663	1.577	14.348	(8.080, 43.037)	1.221
	exon_UTR_ST	51.613	(26.113, 256.666)	2.325	1.790	1.676	31.619	(12.612, NaN)	**1.469**
	exon_UTR_AS	39.068	(22.902, 85.847)	2.058	1.673	1.585	14.791	(6.622, 59.368)	1.231
	intergenic	69.865	(58.361, 85.390)	2.744	1.930	1.777	15.765	(11.084, 27.315)	1.251
B.	paired vs. unpaired								
	intron_UTR_ST	11.099	(4.690, 63.568)	1.340	1.205	1.176	6.153	(1.301, NaN)	0.942
	intron_UTR_AS	9.729	(5.126, 30.131)	1.281	1.158	1.133	NaN	(NaN, NaN)	NaN
	intron_nUTR_ST	11.368	(6.500, 31.849)	**1.351**	**1.213**	**1.184**	4.243	(1.955, 14.564)	**0.811**
	intron_nUTR_AS	8.456	(5.459, 15.317)	1.220	1.108	1.087	4.874	(1.864, 22.898)	0.861
	exon_nUTR_ST	3.131	(2.139, 5.194)	**0.824**	**0.737**	**0.735**	4.303	(2.627, 7.853)	0.816
	exon_nUTR_AS	4.598	(2.971, 8.900)	0.977	0.886	0.877	4.322	(2.236, 12.566)	0.818
	exon_UTR_ST	8.069	(3.866, 39.268)	1.200	1.091	1.071	8.811	(2.830, NaN)	**1.063**
	exon_UTR_AS	6.459	(3.601, 14.835)	1.109	1.010	0.995	4.688	(1.777, 22.219)	0.847
	intergenic	7.710	(6.344, 9.450)	1.181	1.075	1.056	6.102	(4.117, 10.512)	0.939

TABLE B.S5. **Ratio of N_{exp} to N_{obs}, 95% confidence intervals and scaled selection coefficients for drosophilid and hominid RNA folds in various genomic regions.** NOTE.– Min and Max values of $N_e s$ are given in bold letters.

Influence of RNA Structural Features on Selection Coefficients

To investigate the variation of $N_e s$ and its dependence on structural features of the RNA molecule we measured the average size (which is positively correlated with the average distance between pairing nucleotides, Kendall's $\tau = 0.669$, $P < 2.2 \times 10^{-16}$), average helix length and average GC content of all ncRNA molecules in windows of 1-Mb along the *D.melanogaster* chromosomes. Subsequently, these three variables were tested for their combined influence on $N_e s$ using a generalized additive model (GAM) (Wood, 2006), which allows us to determine nonlinear humped relationships between regressors and response (fig. B.S2 and page 85). General trends obtained by the additive model agree with our previous results (Piskol and Stephan, 2008)(fig. B.S2 solid lines). They suggest a negative influence of average helix length on $N_e s$ (for helix length ≥ 5), which is consistent with our previous results that longer helices contain more covariations, wobbles (GU pairs) and mismatches (Parsch et al., 2000; Piskol and Stephan, 2008). A negative relation can also be observed for the GC content and suggests that RNA molecules of higher GC content experience relaxed evolutionary constraints. This observation agrees with the slightly higher GC content of folds in coding regions (exon_nUTR_ST/AS in Table 3.1 on page 30) and their marginally decreased constraints (Table 3, main text). It also points to a different relationship between divergence and GC content in RNA molecules and neutral sites. While the interaction between GC content and divergence shows a U-shape at neutral sites (Előry et al., 2010), our results suggest that divergence increases for higher GC content in folded RNA molecules. On the other hand, the size of the RNA molecule has a positive effect on $N_e s$ leading to higher constraints in larger molecules. This observation can be explained by the retarding effect of recombination on pairing positions that are separated by a greater distance in sequence (Stephan, 1996). It is also important to note that the interplay between the three variables is relevant for the calculation of their effect on $N_e s$ as can be seen from the difference between the GAM estimates that were taking all factors simultaneously into account (solid line) and locally weighted regression (loess) that was performed on each of the factors one by one (dashed line).

Appendix B.

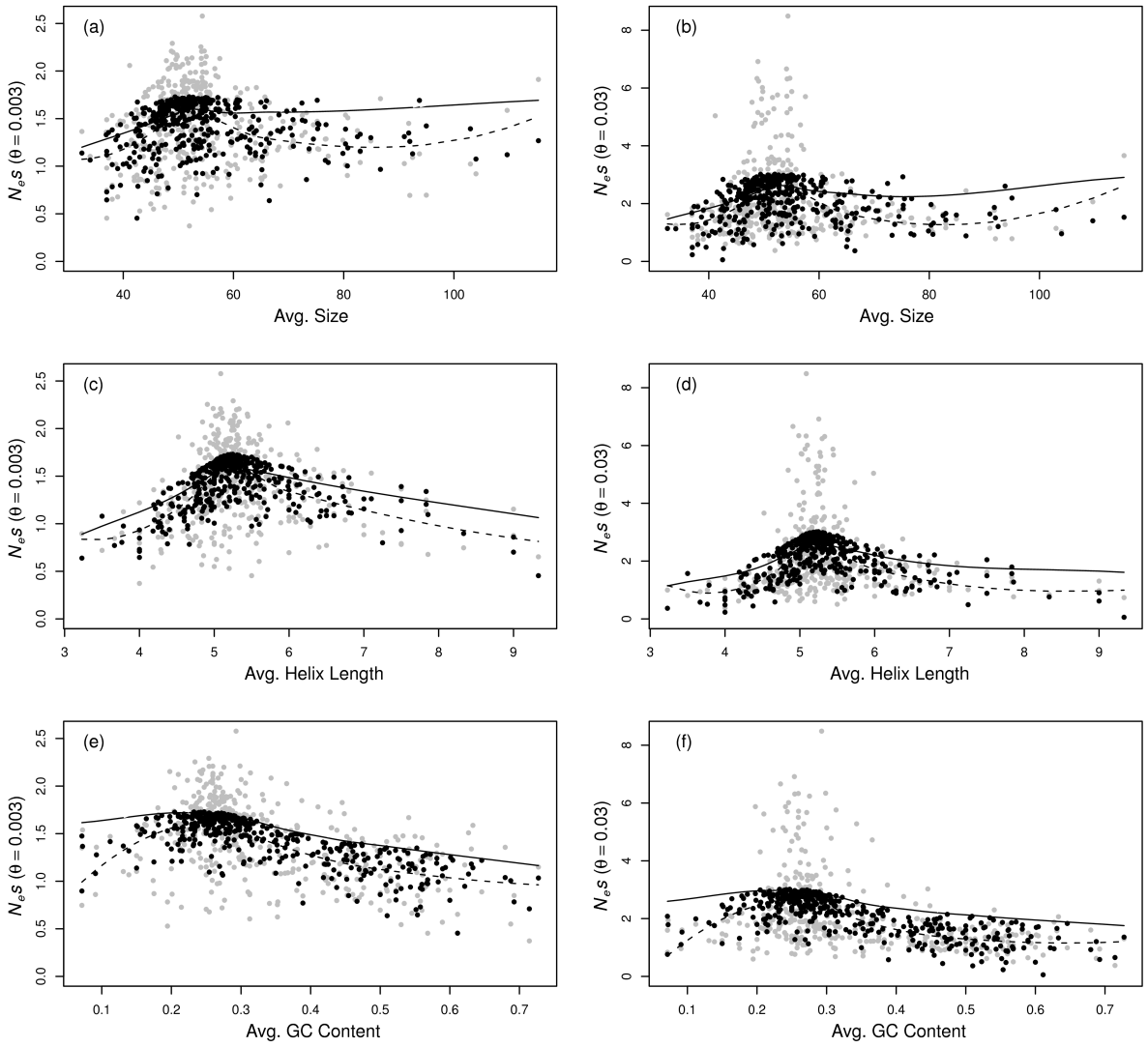

FIGURE B.S2. **Influence of average size, helix length and GC content on the estimated scaled selection coefficients in drosophilid ncRNAs.** Gray dots represent true values of $N_e s$ plotted against avg. size (fig. B.S2a,b), avg. helix length (fig. B.S2c,d) and avg. GC content (fig. B.S2e,f), which were measured for folded molecules in windows of 1-Mb along the chromosomes. Black dots show fitted values of $N_e s$. They were obtained by fitting a generalized additive model (GAM) to the data that takes all three variables simultaneously into account. The black solid line represents values of $N_e s$ predicted by the GAM as a function of the respective variable, holding the remaining two variables at their mean. Dashed lines represent locally weighted regressions for each parameter independently. Plots in the left and right columns show values of $N_e s$ for $\theta = 0.003$ and $\theta = 0.03$, respectively. Note the difference between y-scales for $\theta = 0.003$ and $\theta = 0.03$.

Generalized Additive Models

Generalized additive models (Wood, 2006) extend generalized linear models by replacing a linear predictor of the form $\alpha + \sum_j X_j\beta_j$ with the additive form $\alpha + \sum_j f_j(X_j)$ and linking a response variable Y with mean $\mu = E(Y|X_1, \cdots, X_p)$ to this predictor through

$$g(\mu) = \alpha + \sum_{j=1}^{p} f_j(X_j). \tag{B.1}$$

Thereby the $f_j(\cdot)$ describe arbitrary univariate functions for each predictor variable. Such a model allows us to specify flexible relations between the response variable and its regressors X_j without prior knowledge about the underlying relationship between them.

When applying this framework to our drosophilid data we defined the scaled selection coefficient $N_e s$ for RNA molecules as the response variable and investigated its dependence on three basic characteristics of the RNA structures – namely average size, average helix length, and GC content. This relationship was described by

$$\mathbb{E}[N_e s] = f_1(avg.\,size) + f_2(avg.\,helix\,length) + f_3(GC content). \tag{B.2}$$

Thereby, we chose the identity link function ($g(\mu) = \mu$) and used locally-weighted smoothers (loess) for $f_i, i \in (1, 2, 3)$ with a span of 0.75 (Hastie and Tibshirani, 1990). Furthermore, we made sure that the three variables did not display any interactions by testing the model

$$\mathbb{E}[N_e s] = f_1(avg.\,size) * f_2(avg.\,helix\,length) * f_3(GC content), \tag{B.3}$$

which did not perform significantly better and also made sure that a model using a multivariate function, i.e.,

$$\mathbb{E}[N_e s] = f_1(avg.\,size,\ avg.\,helix\,length,\ GC content) \tag{B.4}$$

did not perform significantly better.

Our results show that all three predictor variables have a significant influence on $N_e s$, which is clearly non-linear (fig. B.S2). The degrees of freedom for each term and P-values for the nonparametric effects are given in Table B.S6 for two different values of θ. A visual inspection of residual plots and QQ plots suggests that the models fit reasonably well (fig. B.S3).

	$\theta = 0.003$		$\theta = 0.03$	
Independent variable	Npar Df	Pr(> |F|)	Npar Df	Pr(> |F|)
Avg. Size	2.5	1.115×10^{-4}***	2.5	9.872×10^{-5}***
Avg. Helix Length	2.8	1.998×10^{-15}***	2.8	2.983×10^{-8}***
GC Content	1.9	1.018×10^{-5}***	1.9	2.283×10^{-4}***

TABLE B.S6. **Degrees of freedom and P-values for nonparametric effects of the generalized additive model from equation (B.2).**

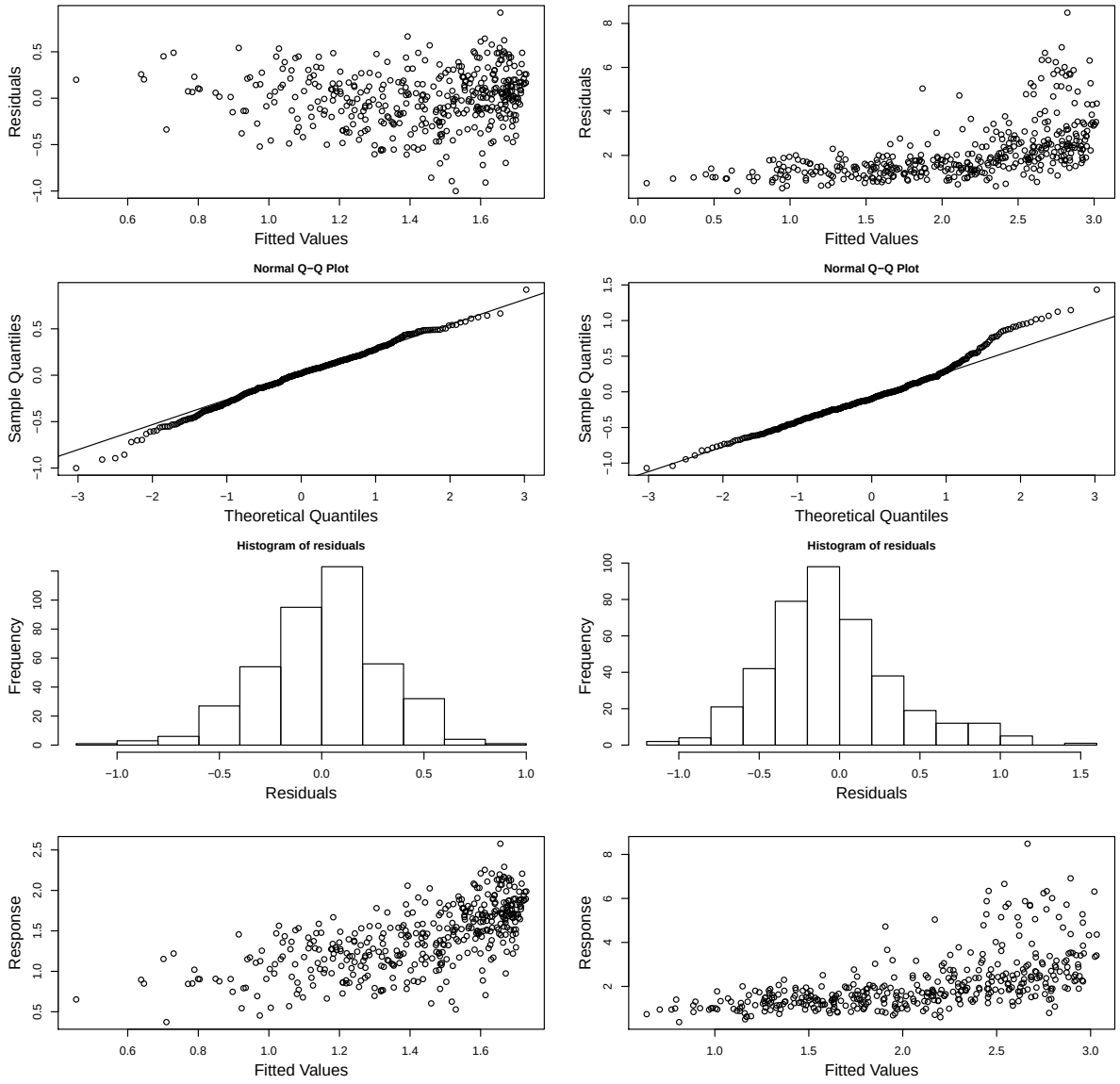

FIGURE B.S3. **Plots of residuals vs. fitted values, normal Q-Q plot, histogram of residuals and plot of response vs. fitted values for $\theta = 0.003$ and $\theta = 0.03$ in the left and right columns, respectively.**

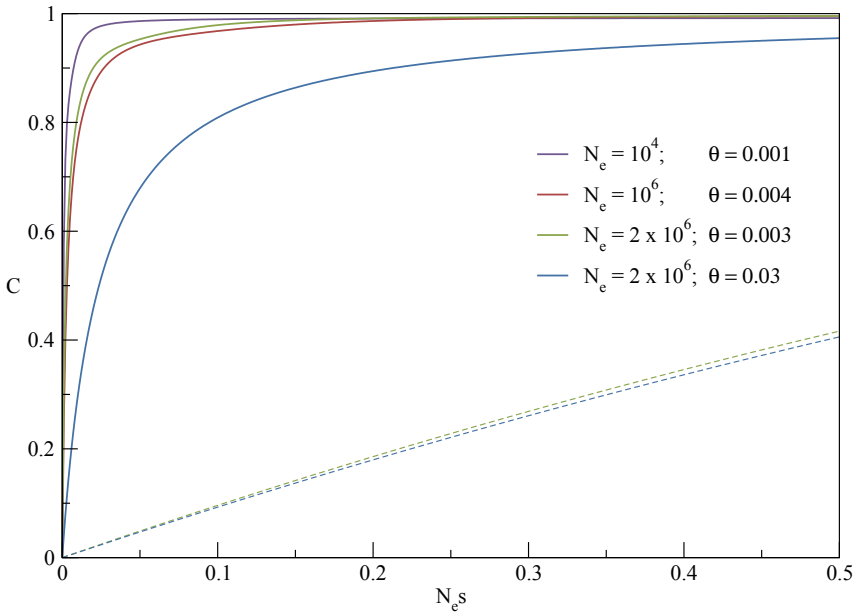

FIGURE C.S1. **Expected selective constraints at coevolving sites (C_{coev}) (solid lines) and independently evolving sites (C_{ind}) (dashed lines) as a function of the scaled selection coefficient $N_e s$.** N_e and θ were set to approximately match the parameters for hominids ($\theta = 0.001$), murids ($\theta = 0.004$) and drosophilids ($\theta = 0.003; \theta = 0.03$), respectively. The trajectories for C_{ind} and C_{coev} were obtained from Kimura's unidirectional models for the expected fixation times of mutant alleles in a population (see main text).

Appendix C.

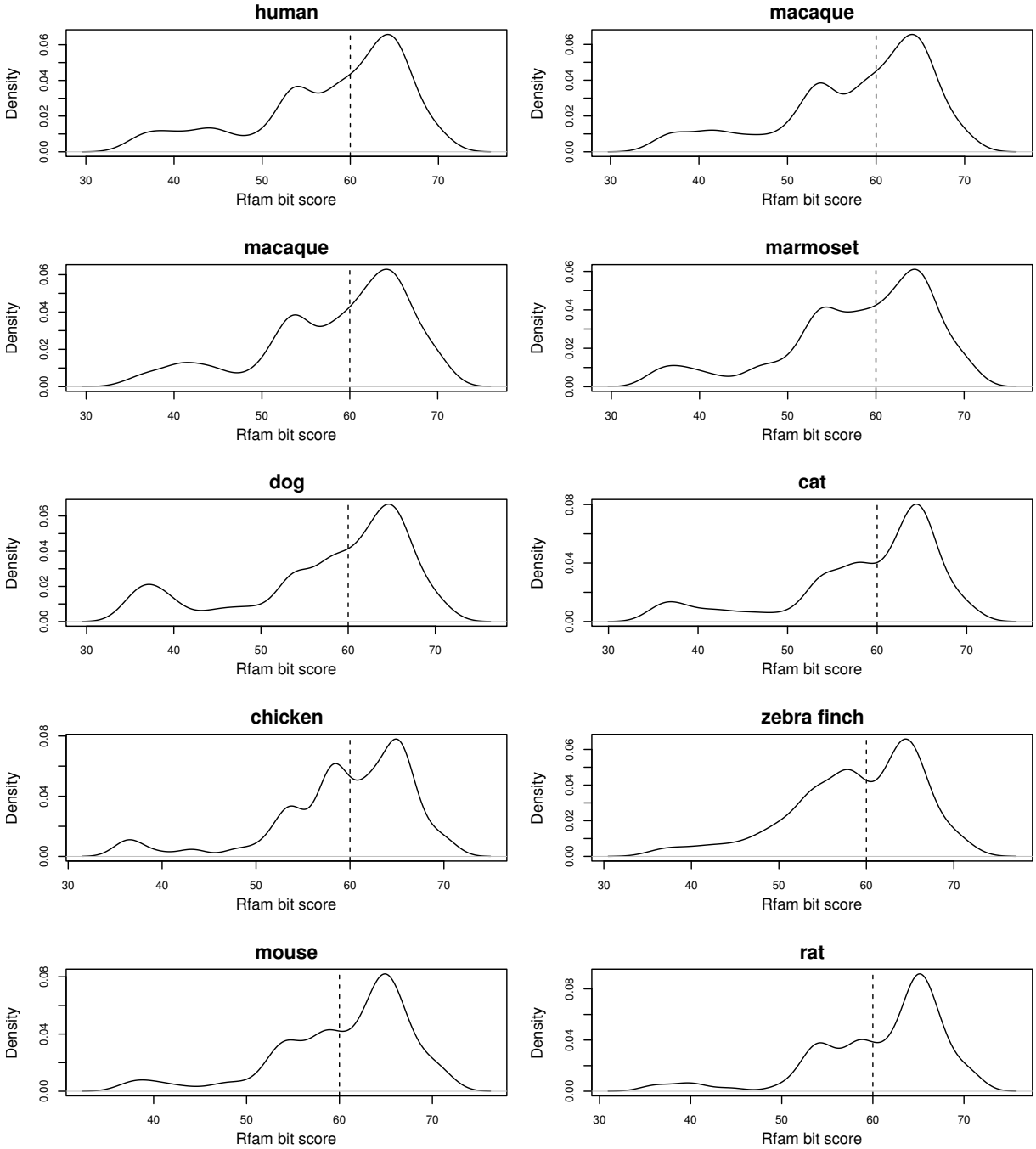

FIGURE C.S2. **Densities of INFERNAL bit scores in orthologous tRNAs for all analyzed vertebrate species pairs.** The densities of scores for the two species in a pair are shown next to each other. The vertical dashed lines represent the score ($S = 60$) at which the data was split into the two sets of low and high similarity to the Rfam covariance model.

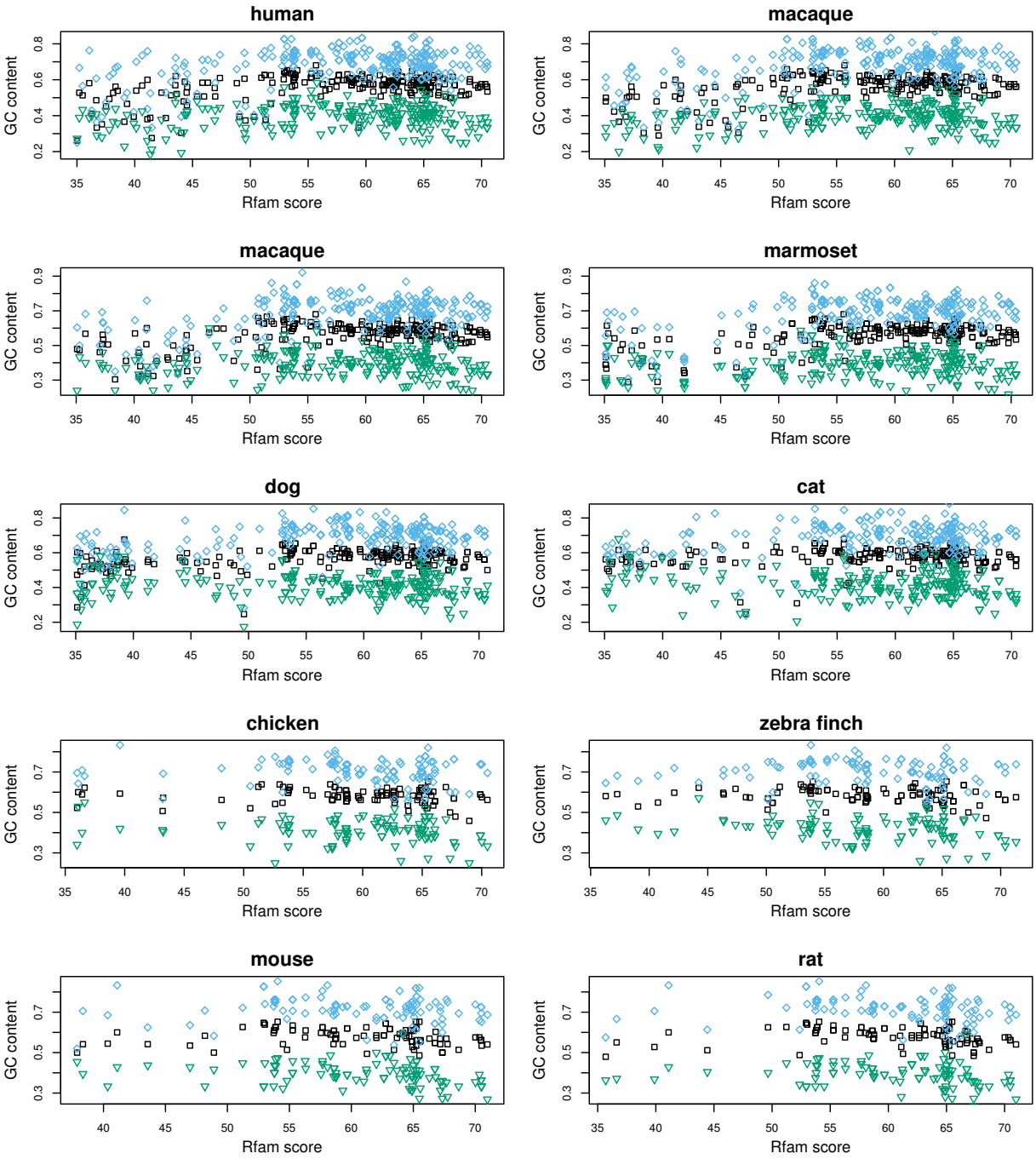

FIGURE C.S3. **GC content in dependency of INFERNAL bit scores in orthologous tRNAs for all analyzed vertebrate species pairs.** The GC contents for the two species in a pair are shown next to each other. The GC content at all positions is represented by black squares. The GC content at paired and unpaired positions is symbolized by blue diamonds and green triangles, respectively.

Appendix C.

	Species pair	C_{paired} (95% CI)	$C_{unpaired}$ (95% CI)	$\lvert C_{paired} - C_{unpaired} \rvert$
A.	$score > 35$			
	human/macaque	0.884 (0.839, 0.933)	0.698 (0.623, 0.779)	0.186
	macaque/marmoset	0.899 (0.864, 0.932)	0.787 (0.737, 0.839)	0.112
	dog/cat	0.911 (0.885, 0.940)	0.848 (0.812, 0.883)	0.063
	chicken/zebra finch	0.953 (0.942, 0.966)	0.917 (0.897, 0.939)	0.036
	mouse/rat	0.989 (0.979, 1.000)	0.946 (0.914, 0.979)	0.043
	D.mel/D.yak	0.996 (0.994, 0.999)	0.982 (0.975, 0.989)	0.014
	D.mel/D.sim	0.984 (0.976, 0.992)	0.976 (0.962, 0.989)	0.008
B.	$35 < score < 60$			
	human/macaque	0.814 (0.736, 0.896)	0.536 (0.406, 0.663)	0.278
	macaque/marmoset	0.820 (0.762, 0.887)	0.633 (0.545, 0.723)	0.187
	dog/cat	0.833 (0.785, 0.889)	0.773 (0.712, 0.835)	0.060
	chicken/zebra finch	0.936 (0.918, 0.956)	0.892 (0.862, 0.925)	0.044
	mouse/rat	0.987 (0.974, 1.000)	0.902 (0.842, 0.971)	0.085
	D.mel/D.yak	0.991 (0.986, 0.997)	0.962 (0.946, 0.979)	0.029
	D.mel/D.sim	0.959 (0.940, 0.981)	0.937 (0.902, 0.975)	0.022
C.	$score \geq 60$			
	human/macaque	0.959 (0.928, 1.000)	0.891 (0.830, 0.965)	0.068
	macaque/marmoset	0.976 (0.957, 1.000)	0.957 (0.928, 0.995)	0.019
	dog/cat	0.977 (0.961, 0.997)	0.925 (0.891, 0.958)	0.052
	chicken/zebra finch	0.973 (0.961, 0.988)	0.951 (0.928, 0.976)	0.022
	mouse/rat	0.991 (0.982, 1.000)	0.980 (0.961, 1.000)	0.011
	D.mel/D.yak	0.999 (0.999, 1.000)	0.994 (0.989, 0.999)	0.005
	D.mel/D.sim	0.998 (0.997, 1.000)	1.000 (1.000, 1.000)	0.002

TABLE C.S1. **Selective constraints for paired (C_{paired}) and unpaired (C_{unpaired}) regions based on a MLOCARNA alignment of tRNAs in various species pairs.** The difference $|C_{\text{paired}} - C_{\text{unpaired}}|$ is calculated for the point estimates.

	Species pair	C_{paired} (95% CI)	$C_{unpaired}$ (95% CI)	$\lvert C_{paired} - C_{unpaired} \rvert$
A.	$score > 35$			
	human/macaque	0.902 (0.857, 0.954)	0.742 (0.679, 0.813)	0.160
	macaque/marmoset	0.905 (0.870, 0.944)	0.826 (0.783, 0.873)	0.079
	dog/cat	0.920 (0.892, 0.949)	0.865 (0.836, 0.895)	0.055
	chicken/zebra finch	0.956 (0.944, 0.967)	0.923 (0.905, 0.941)	0.033
	mouse/rat	0.994 (0.988, 1.000)	0.944 (0.915, 0.977)	0.050
	D.mel/D.yak	0.998 (0.996, 1.000)	0.983 (0.977, 0.990)	0.015
	D.mel/D.sim	0.989 (0.982, 0.997)	0.970 (0.955, 0.985)	0.019
B.	$35 < score < 60$			
	human/macaque	0.815 (0.735, 0.907)	0.627 (0.525, 0.733)	0.188
	macaque/marmoset	0.818 (0.748, 0.889)	0.707 (0.639, 0.778)	0.111
	dog/cat	0.869 (0.821, 0.927)	0.792 (0.742, 0.847)	0.077
	chicken/zebra finch	0.942 (0.925, 0.962)	0.897 (0.866, 0.927)	0.045
	mouse/rat	0.986 (0.971, 1.000)	0.907 (0.847, 0.969)	0.079
	D.mel/D.yak	0.994 (0.990, 1.000)	0.964 (0.951, 0.980)	0.030
	D.mel/D.sim	0.969 (0.950, 0.994)	0.931 (0.898, 0.967)	0.038
C.	$score \geq 60$			
	human/macaque	0.989 (0.977, 1.000)	0.886 (0.824, 0.961)	0.103
	macaque/marmoset	0.986 (0.973, 1.000)	0.963 (0.938, 0.994)	0.023
	dog/cat	0.962 (0.939, 0.987)	0.939 (0.909, 0.971)	0.023
	chicken/zebra finch	0.972 (0.958, 0.987)	0.957 (0.938, 0.978)	0.015
	mouse/rat	1.000 (1.000, 1.000)	0.974 (0.949, 1.000)	0.026
	D.mel/D.yak	0.999 (0.997, 1.000)	0.995 (0.991, 1.000)	0.004
	D.mel/D.yak	1.000 (1.000, 1.000)	0.997 (0.994, 1.000)	0.003

TABLE C.S2. **Selective constraints for paired ($C_{\mathbf{paired}}$) and unpaired ($C_{\mathbf{unpaired}}$) regions based on a MUSCLE alignment of tRNAs in various species pairs.** The difference $|C_{\text{paired}} - C_{\text{unpaired}}|$ is calculated for the point estimates.

Appendix C.

	Species pair	C_{paired} (95% CI)	$C_{unpaired}$ (95% CI)	$\lvert C_{paired} - C_{unpaired} \rvert$
A.	$score > 35$			
	human/macaque	0.730 (0.645, 0.826)	0.589 (0.511, 0.673)	0.141
	macaque/marmoset	0.824 (0.775, 0.879)	0.736 (0.682, 0.790)	0.088
	dog/cat	0.873 (0.837, 0.912)	0.796 (0.759, 0.838)	0.077
	chicken/zebra finch	0.896 (0.875, 0.922)	0.861 (0.836, 0.888)	0.035
	mouse/rat	0.961 (0.929, 1.000)	0.947 (0.920, 0.978)	0.014
	D.mel/D.yak	0.989 (0.986, 0.993)	0.927 (0.914, 0.941)	0.062
	D.mel/D.sim	0.969 (0.958, 0.984)	0.889 (0.863, 0.918)	0.080
B.	$35 < score < 60$			
	human/macaque	0.512 (0.367, 0.670)	0.405 (0.274, 0.547)	0.107
	macaque/marmoset	0.662 (0.562, 0.769)	0.523 (0.431, 0.632)	0.139
	dog/cat	0.778 (0.709, 0.859)	0.615 (0.539, 0.696)	0.163
	chicken/zebra finch	0.862 (0.826, 0.898)	0.801 (0.762, 0.842)	0.061
	mouse/rat	0.908 (0.832, 1.000)	0.906 (0.854, 0.972)	0.002
	D.mel/D.yak	0.970 (0.959, 0.982)	0.816 (0.786, 0.850)	0.154
	D.mel/D.sim	0.916 (0.885, 0.954)	0.724 (0.650, 0.792)	0.192
C.	$score \geq 60$			
	human/macaque	0.959 (0.919, 1.000)	0.786 (0.702, 0.884)	0.173
	macaque/marmoset	0.977 (0.953, 1.000)	0.942 (0.908, 0.970)	0.035
	dog/cat	0.958 (0.931, 0.990)	0.948 (0.925, 0.975)	0.010
	chicken/zebra finch	0.936 (0.912, 0.961)	0.940 (0.917, 0.967)	0.004
	mouse/rat	1.000 (1.000, 1.000)	0.977 (0.954, 1.000)	0.023
	D.mel/D.yak	0.999 (0.997, 1.000)	0.995 (0.991, 1.000)	0.004
	D.mel/D.sim	1.000 (1.000, 1.000)	0.997 (0.994, 1.000)	0.003

TABLE C.S3. **Selective constraints for paired (C_{paired}) and unpaired (C_{unpaired}) regions based on an INFERNAL alignment of tRNAs in various species pairs.** The difference $|C_{\text{paired}} - C_{\text{unpaired}}|$ is calculated for the point estimates.

Species pair	GC%	Divergence	# Sites	$k_{AT\leftrightarrow TA}$	$k_{CG\leftrightarrow GC}$	$k_{AT\rightarrow GC}$	$k_{GC\rightarrow AT}$
human/macaque	0.4503	0.0645	63,996	0.0078	0.0148	0.0377	0.0636
macaque/marmoset	0.4685	0.1157	108,081,432	0.0166	0.0240	0.0628	0.1058
dog/cat	0.4442	0.1902	89,665,142	0.0299	0.0406	0.0960	0.1598
chicken/zebra finch	0.4323	0.3574	3,204,089	0.0758	0.0729	0.1524	0.2452
mouse/rat	0.4914	0.1836	39,232,282	0.0336	0.0355	0.0899	0.1521
D.mel/D.yak[a]	0.3406	0.3016	80,225	0.0995	0.0703	0.1340	0.1983
D.mel/D.yak[b]	0.3359	0.2997	73,952	0.0998	0.0702	0.1334	0.1965
D.mel/D.yak[c]	0.3947	0.3249	6,273	0.0946	0.0720	0.1405	0.2175
D.mel/D.sim	0.3422	0.1064	81,406	0.0330	0.0235	0.0563	0.0905

TABLE C.S4. **Nucleotide content and substitution rates in sequences used as neutral standards.** Divergence values are given after applying the Jukes-Cantor correction for multiple hits. Values for vertebrates were obtained from ancestral repeats, while in Drosophila positions 8-30 of short introns ($\leq$65nt) were analyzed. [a]all chromosomes; [b]autosomes; [c]X chromosome.

Appendix C.

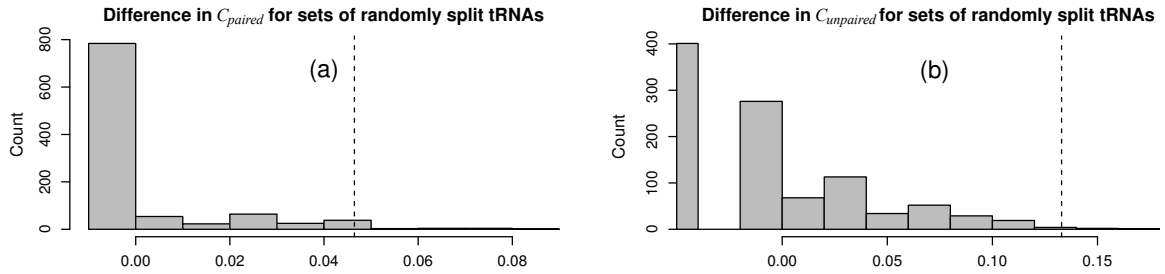

FIGURE C.S4. **Histogram of differences in C at (a) paired and (b) unpaired positions between sets of 90 and 5 tRNAs that were created by randomly splitting 95 orthologous tRNAs of *D. melanogaster* and *D. yakuba* from the peripheral set 1000 times.**

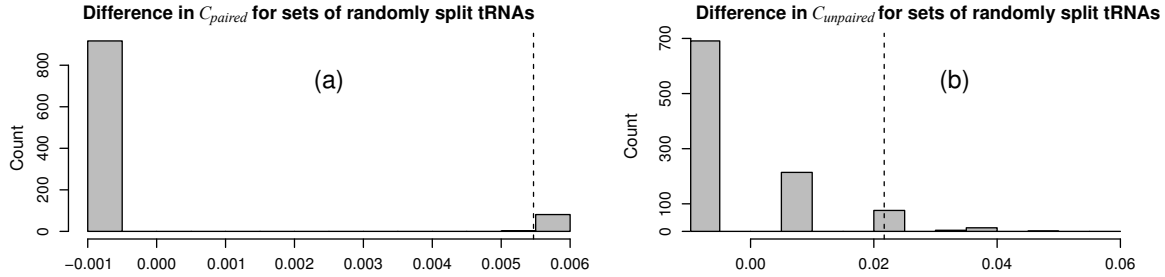

FIGURE C.S5. **Histogram of differences in C at (a) paired and (b) unpaired positions between sets of 166 and 16 tRNAs that were created by randomly splitting 182 orthologous tRNAs of *D. melanogaster* and *D. yakuba* from the core set 1000 times.**

Species pair	C_{paired} (95% CI)	$C_{unpaired}$ (95% CI)	$\lvert C_{paired} - C_{unpaired} \rvert$
human/macaque	0.759 (0.709, 0.808)	0.595 (0.523, 0.675)	0.164
mouse/rat	0.848 (0.818, 0.880)	0.805 (0.761, 0.849)	0.043
D.mel/D.yak	0.969 (0.958, 0.983)	0.865 (0.824, 0.905)	0.104

TABLE C.S5. **Selective constraints for paired ($C_{\mathbf{paired}}$) and unpaired ($C_{\mathbf{unpaired}}$) regions of hominid, murid and drosophilid miRNAs.** The difference $|C_{\text{paired}} - C_{\text{unpaired}}|$ is calculated for the point estimates.

Bibliography

1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. Nature. 467(7319):1061–1073.

Akashi H. 1994. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. Genetics. 136(3):927–935.

Amaral PP, Dinger ME, Mercer TR, Mattick JS. 2008. The eukaryotic genome as an RNA machine. Science. 319(5871):1787–1789.

Anant S, Davidson NO. 2001. Molecular mechanisms of apolipoprotein B mRNA editing. Curr Opin Lipidol. 12(2):159–165.

Ancel LW, Fontana W. 2000. Plasticity, evolvability, and modularity in RNA. J Exp Zool. 288(3):242–283.

Andolfatto P. 2001. Contrasting patterns of X-linked and autosomal nucleotide variation in *Drosophila melanogaster* and *Drosophila simulans*. Mol Biol Evol. 18(3):279–290.

Andolfatto P, Wong KM, Bachtrog D. 2011. Effective population size and the efficacy of selection on the X chromosomes of two closely related Drosophila species. Genome Biol Evol. 3:114–128.

Atwell S, et al. 2010. Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature. 465(7298):627–631.

Baines JF, Harr B. 2007. Reduced X-linked diversity in derived populations of house mice. Genetics. 175(4):1911–1921.

Baines JF, Parsch J, Stephan W. 2004. Pleiotropic effect of disrupting a conserved sequence involved in a long-range compensatory interaction in the Drosophila *Adh* gene. Genetics. 166(1):237–242.

Bass BL. 2002. RNA editing by adenosine deaminases that act on RNA. Annu Rev Biochem. 71:817–846.

Begun DJ, et al. 2007. Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. PLoS Biol. 5(11):e310.

Bernardi G. 2000. Isochores and the evolutionary genomics of vertebrates. Gene. 241(1):3–17.

Betancourt AJ, Presgraves DC, Swanson WJ. 2002. A test for faster X evolution in Drosophila. Mol Biol Evol. 19(10):1816–1819.

Blanc V, Davidson NO. 2003. C-to-U RNA editing: mechanisms leading to genetic diversity. J Biol Chem. 278(3):1395–1398.

Bradley RK, et al. 2009. Evolutionary modeling and prediction of non-coding RNAs in Drosophila. PLoS One. 4(8):e6478.

Brown KM, et al. 2010. Compensatory mutations restore fitness during the evolution of dihydrofolate reductase. Mol Biol Evol. 27(12):2682–2690.

Brusa R, et al. 1995. Early-onset epilepsy and postnatal lethality associated with an editing-deficient GluR-B allele in mice. Science. 270(5242):1677–1680.

Bullock SL, Zicha D, Ish-Horowicz D. 2003. The Drosophila hairy RNA localization signal modulates the kinetics of cytoplasmic mRNA transport. EMBO J. 22(10):2484–2494.

Cao F, et al. 2009. Dicer independent small RNAs associate with telomeric heterochromatin. RNA. 15(7):1274–1281.

Chamary JV, Parmley JL, Hurst LD. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet. 7(2):98–108.

Charlesworth B. 2009. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. Nat Rev Genet. 10(3):195–205.

Chen Y, et al. 1999. RNA secondary structure and compensatory evolution. Genes Genet Syst. 74(6):271–286.

Chen Y, Stephan W. 2003. Compensatory evolution of a precursor messenger RNA secondary structure in the *Drosophila melanogaster Adh* gene. Proc Natl Acad Sci USA. 100(20):11499–11504.

Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature. 437(7055):69–87.

Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R. 2005. Ascertainment bias in studies of human genome-wide polymorphism. Genome Res. 15(11):1496–1502.

Clément Y, Arndt PF. 2011. Substitution patterns are under different influences in primates and rodents. Genome Biol Evol. doi: 10.1093/gbe/evr011.

Connallon T. 2007. Adaptive protein evolution of X-linked and autosomal genes in Drosophila: implications for faster-X hypotheses. Mol Biol Evol. 24(11):2566–2572.

Cowperthwaite MC, Economo EP, Harcombe WR, Miller EL, Meyers LA. 2008. The ascent of the abundant: how mutational networks constrain evolution. PLoS Comput Biol. 4(7):e1000110.

Crick F. 1970. Central dogma of molecular biology. Nature. 227(5258):561–563.

Draghi JA, Parsons TL, Wagner GP, Plotkin JB. 2010. Mutational robustness can facilitate adaptation. Nature. 463(7279):353–355.

Drosophila 12 Genomes Consortium. 2007. Evolution of genes and genomes on the Drosophila phylogeny. Nature. 450:203–219.

Duret L, Arndt PF. 2008. The impact of recombination on nucleotide substitutions in the human genome. PLoS Genet. 4(5):e1000071.

Duret L, Chureau C, Samain S, Weissenbach J, Avner P. 2006. The *Xist* RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. Science. 312(5780):1653–1655.

Dutheil JY, Jossinet F, Westhof E. 2010. Base pairing constraints drive structural epistasis in ribosomal RNA sequences. Mol Biol Evol. 27(8):1868–1876.

Eddy SR, Durbin R. 1994. RNA sequence analysis using covariance models. Nucleic Acids Res. 22(11):2079–2088.

Edgar RC. 2004. Muscle: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 5:113.

Előry 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(1):177–192.

Elena SF, Sanjuán R. 2008. The effect of genetic robustness on evolvability in digital organisms. BMC Evol Biol. 8:284.

Evans BJ, Pin L, Melnick DJ, Wright SI. 2010. Sex-linked inheritance in macaque monkeys: implications for effective population size and dispersal to Sulawesi. Genetics. 185(3):923–937.

Eyre-Walker A, Hurst LD. 2001. The evolution of isochores. Nat Rev Genet. 2(7):549–555.

Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. Nat Rev Genet. 8(8):610–618.

Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol Biol Evol. 26(9):2097–2108.

Eyre-Walker A, Keightley PD, Smith NGC, Gaffney D. 2002. Quantifying the slightly deleterious mutation model of molecular evolution. Mol Biol Evol. 19(12):2142–2149.

Eyre-Walker A, Woolfit M, Phelps T. 2006. The distribution of fitness effects of new deleterious amino acid mutations in humans. Genetics. 173(2):891–900.

Felsenstein J. 1984. Distance methods for inferring phylogenies: A justification. Evolution. 38:16–24.

Gaffney DJ, Keightley PD. 2008. Effect of the assignment of ancestral CpG state on the estimation of nucleotide substitution rates in mammals. BMC Evol Biol. 8:265.

Galasso M, Sana ME, Volinia S. 2010. Non-coding RNAs: a key to future personalized molecular therapy? Genome Med. 2(2):12.

Gardner PP, et al. 2009. Rfam: updates to the RNA families database. Nucleic Acids Res. 37(Database issue):D136–D140.

Gillespie JH. 1999. The role of population size in molecular evolution. Theor Popul Biol. 55(2):145–156.

Gossmann TI, Song BH, Windsor AJ, Mitchell-Olds T, Dixon CJ, Kapralov MV, Filatov DA, Eyre-Walker A. 2010. Genome wide analyses reveal little evidence for adaptive evolution in many plant species. Mol Biol Evol. 27(8):1822–1832.

Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. 2008. miRBase: tools for microRNA genomics. Nucleic Acids Res. 36(Database issue):D154–D158.

Grohmann M, et al. 2010. Alternative splicing and extensive RNA editing of human TPH2 transcripts. PLoS One. 5(1):e8956.

Gutierrez-Aguilar AL, Piskol R, Beitzinger M, Zhu JY, Kruspe D, Aszodi A, Moser M, Englert C, Meister G. 2010. The small RNA expression profile of the developing murine urinary and reproductive systems. FEBS Lett. 584(21):4426–4434.

Halligan DL, Eyre-Walker A, Andolfatto P, Keightley PD. 2004. Patterns of evolutionary constraints in intronic and intergenic DNA of Drosophila. Genome Res. 14(2):273–279.

Hastie T, Tibshirani R. 1990. Generalized additive models. Chapman & Hall/CRC.

Hersberger M, Patarroyo-White S, Arnold KS, Innerarity TL. 1999. Phylogenetic analysis of the apolipoprotein B mRNA-editing region. evidence for a secondary structure between the mooring sequence and the 3' efficiency element. J Biol Chem. 274(49):34590–34597.

Higgs PG. 1998. Compensatory neutral mutations and the evolution of RNA. Genetica. 102-103(1-6):91–101.

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR. 2005. Whole-genome patterns of common DNA variation in three human populations. Science. 307(5712):1072–1079.

Hobert O. 2008. Gene regulation by transcription factors and microRNAs. Science. 319(5871):1785–1786.

Höchsmann M, Voss B, Giegerich R. 2004. Pure multiple RNA secondary structure alignments: a progressive profile approach. IEEE/ACM Trans Comput Biol Bioinform. 1(1):53–62.

Hoede C, Denamur E, Tenaillon O. 2006. Selection acts on DNA secondary structures to decrease transcriptional mutagenesis. PLoS Genet. 2(11):e176.

Hofacker IL, et al. 1994. Fast folding and comparison of RNA secondary structures. Monatshefte für Chemie. 125:167–188.

Hofacker IL, Fekete M, Stadler PF. 2002. Secondary structure prediction for aligned RNA sequences. J Mol Biol. 319(5):1059–1066.

Holmes I. 2005. Accelerated probabilistic inference of RNA structure evolution. BMC Bioinformatics. 6:73.

Hoopengardner B, Bhalla T, Staber C, Reenan R. 2003. Nervous system targets of RNA editing identified by comparative genomics. Science. 301(5634):832–836.

Howe KJ, Ares M. 1997. Intron self-complementarity enforces exon inclusion in a yeast pre-mRNA. Proc Natl Acad Sci U S A. 94(23):12467–12472.

Hutter S, Li H, Beisswanger S, de Lorenzo D, Stephan W. 2007. Distinctly different sex ratios in African and European populations of *Drosophila melanogaster* inferred from chromosomewide single nucleotide polymorphism data. Genetics. 177(1):469–480.

Ichinose M, Iizuka M, Kado T, Takefu M. 2008. Compensatory evolution in diploid populations. Theor Popul Biol. 74(2):199–207.

Innan H, Stephan W. 2001. Selection intensity against deleterious mutations in RNA secondary structures and rate of compensatory nucleotide substitutions. Genetics. 159(1):389–399.

International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature. 449(7164):851–861.

Irion U, Johnston DS. 2007. *bicoid* RNA localization requires specific binding of an endosomal sorting complex. Nature. 445(7127):554–558.

Iwasa Y, Michor F, Nowak MA. 2004. Stochastic tunnels in evolutionary dynamics. Genetics. 166(3):1571–1579.

Jennings WB, Edwards SV. 2005. Speciational history of Australian grass finches (Poephila) inferred from thirty gene trees. Evolution. 59(9):2033–2047.

Jurka J, et al. 2005. Repbase update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 110(1-4):462–467.

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(14):3059–3066.

Keightley PD, Eyre-Walker A. 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics. 177(4):2251–2261.

Keightley PD, Eyre-Walker A. 2010. What can we learn about the distribution of fitness effects of new mutations from DNA sequence data? Phil. Trans. R. Soc. B. 365(1544):1187–1193.

Keightley PD, Gaffney DJ. 2003. Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. Proc Natl Acad Sci U S A. 100(23):13402–13406.

Kent WJ, et al. 2002. The human genome browser at UCSC. Genome Res. 12(6):996–1006.

Khelifi A, Meunier J, Duret L, Mouchiroud D. 2006. GC content evolution of the human and mouse genomes: insights from the study of processed pseudogenes in regions of different recombination rates. J Mol Evol. 62(6):745–752.

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

Kimura M. 1980. Average time until fixation of a mutant allele in a finite population under continued mutation pressure: Studies by analytical, numerical, and pseudo-sampling methods. Proc Natl Acad Sci U S A. 77(1):522–526.

Kimura M. 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press, England.

Kimura M. 1985. The role of compensatory neutral mutations in molecular evolution. J. Genet. 64:7–19.

Kimura M, Ohta T. 1969. The average number of generations until fixation of a mutant gene in a finite population. Genetics. 61(3):763–771.

Kimura M, Takahata N. 1983. Selective constraint in protein polymorphism: study of the effectively neutral mutation model by using an improved pseudosampling method. Proc Natl Acad Sci U S A. 80(4):1048–1052.

Kirby DA, Muse SV, Stephan W. 1995. Maintenance of pre-mRNA secondary structure by epistatic selection. Proc Natl Acad Sci U S A. 92(20):9047–9051.

Knudsen B, Hein J. 1999. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. Bioinformatics. 15(6):446–454.

Knudsen B, Hein J. 2003. Pfold: RNA secondary structure prediction using stochastic context-free grammars. Nucleic Acids Res. 31(13):3423–3428.

Koonin EV, Wolf YI. 2010. Constraints and plasticity in genome and molecular-phenome evolution. Nat Rev Genet. 11(7):487–498.

Lander ES, et al. 2001. Initial sequencing and analysis of the human genome. Nature. 409(6822):860–921.

Leontis NB, Westhof E. 2001. Geometric nomenclature and classification of RNA base pairs. RNA. 7(4):499–512.

Li H, Stephan W. 2006. Inferring the demographic history and rate of adaptive substitution in *Drosophila*. PLoS Genet. 2(10):e166.

Li JB, et al. 2009. Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. Science. 324(5931):1210–1213.

Lunter G, Ponting CP, Hein J. 2006. Genome-wide identification of human functional dna using a neutral indel model. PLoS Comput Biol. 2(1):e5.

Lynch M. 2010. Scaling expectations for the time to establishment of complex adaptations. Proc Natl Acad Sci U S A. 107(38):16577–16582.

Ma Z, Coruh C, Axtell MJ. 2010. *Arabidopsis lyrata* small RNAs: transient MIRNA and small interfering RNA loci within the Arabidopsis genus. Plant Cell. 22(4):1090–1103.

MacDonald PM. 1990. *bicoid* mRNA localization signal: phylogenetic conservation of function and RNA secondary structure. Development. 110(1):161–171.

Mank JE, Vicoso B, Berlin S, Charlesworth B. 2010. Effective population size and the faster-X effect: empirical results and their interpretation. Evolution. 64(3):663–674.

Mattick JS. 2009. The genetic signatures of noncoding RNAs. PLoS Genet. 5(4):e1000459.

Meer MV, Kondrashov AS, Artzy-Randrup Y, Kondrashov FA. 2010. Compensatory evolution in mitochondrial tRNAs navigates valleys of low fitness. Nature. 464(7286):279–282.

Meister G, Tuschl T. 2004. Mechanisms of gene silencing by double-stranded RNA. Nature. 431(7006):343–349.

Mimouni NK, Lyngsø RB, Griffiths-Jones S, Hein J. 2009. An analysis of structural influences on selection in RNA genes. Mol Biol Evol. 26(1):209–216.

Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature. 420(6915):520–562.

Mural RJ, et al. 2002. A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. Science. 296(5573):1661–1671.

Nagano T, et al. 2008. The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. Science. 322(5908):1717–1720.

Nawrocki EP, Kolbe DL, Eddy SR. 2009. Infernal 1.0: inference of RNA alignments. Bioinformatics. 25(10):1335–1337.

Nei M. 2005. Selectionism and neutralism in molecular evolution. Mol Biol Evol. 22(12):2318–2342.

Nei M, Suzuki Y, Nozawa M. 2010. The neutral theory of molecular evolution in the genomic era. Annu Rev Genomics Hum Genet. 11:265–289.

Nielsen R, Hubisz MJ, Clark AG. 2004. Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. Genetics. 168(4):2373–2382.

Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol. 302(1):205–217.

Ohta T. 1972. Population size and rate of evolution. J Mol Evol. 1(3):305–314.

Ohta T. 1973. Slightly deleterious mutant substitutions in evolution. Nature. 246(5428):96–98.

Ohta T, Gillespie JH. 1996. Development of neutral and nearly neutral theories. Theor Popul Biol. 49(2):128–142.

Pang KC, Frith MC, Mattick JS. 2006. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. Trends Genet. 22(1):1–5.

Parsch J, Braverman JM, Stephan W. 2000. Comparative sequence analysis and patterns of covariation in RNA secondary structures. Genetics. 154(2):909–921.

Parsch J, Novozhilov S, Saminadin-Peter SS, Wong KM, Andolfatto P. 2010. On the utility of short intron sequences as a reference for the detection of positive and negative selection in Drosophila. Mol Biol Evol. 27(6):1226–1234.

Parter M, Kashtan N, Alon U. 2008. Facilitated variation: how evolution learns from past environments to generalize to new environments. PLoS Comput Biol. 4(11):e1000206.

Paz-Yaacov N, et al. 2010. Adenosine-to-inosine RNA editing shapes transcriptome diversity in primates. Proc Natl Acad Sci U S A. 107(27):12174–12179.

Pedersen JS, et al. 2006. Identification and classification of conserved RNA secondary structures in the human genome. PLoS Comput Biol. 2(4):e33.

Piganeau G, Eyre-Walker A. 2009. Evidence for variation in the effective population size of animal mitochondrial DNA. PLoS One. 4(2):e4396.

Piganeau G, Mouchiroud D, Duret L, Gautier C. 2002. Expected relationship between the silent substitution rate and the GC content: implications for the evolution of isochores. J Mol Evol. 54(1):129–133.

Piskol R, Stephan W. 2008. Analyzing the evolution of RNA secondary structures in vertebrate introns using Kimura's model of compensatory fitness interactions. Mol Biol Evol. 25(11):2483–2492.

Piskol R, Stephan W. 2011. Selective constraints in conserved folded RNAs of drosophilid and hominid genomes. Mol Biol Evol. 28(4):1519–1529.

Pitt JN, Ferré-D'Amaré AR. 2010. Rapid construction of empirical RNA fitness landscapes. Science. 330(6002):376–379.

Riedmann EM, Schopoff S, Hartner JC, Jantsch MF. 2008. Specificity of ADAR-mediated RNA editing in newly identified targets. RNA. 14(6):1110–1118.

Rinn JL, et al. 2007. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell. 129(7):1311–1323.

Rogers HH, Bergman CM, Griffiths-Jones S. 2010. The evolution of tRNA genes in Drosophila. Genome Biol Evol. 2:467–477.

Rose D, et al. 2007. Computational RNomics of drosophilids. BMC Genomics. 8:406.

Rousset F, Pélandakis M, Solignac M. 1991. Evolution of compensatory substitutions through G.U intermediate state in Drosophila rRNA. Proc Natl Acad Sci USA. 88(22):10032–10036.

Seemann SE, Gorodkin J, Backofen R. 2008. Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments. Nucleic Acids Res. 36(20):6355–6362.

Shah RR, et al. 1991. Sequence requirements for the editing of apolipoprotein B mRNA. J Biol Chem. 266(25):16301–16304.

Siepel A, Haussler D. 2004. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. Mol Biol Evol. 21(3):468–488.

Silander OK, Tenaillon O, Chao L. 2007. Understanding the evolutionary fate of finite populations: the dynamics of mutational effects. PLoS Biol. 5(4):e94.

Singh ND, Macpherson JM, Jensen JD, Petrov DA. 2007. Similar levels of X-linked and autosomal nucleotide variation in African and non-African populations of *Drosophila melanogaster*. BMC Evol Biol. 7:202.

Slotte T, Foxe JP, Hazzouri KM, Wright SI. 2010. Genome-wide evidence for efficient positive and purifying selection in capsella grandiflora, a plant species with a large effective population size. Mol Biol Evol. 27(8):1813–1821.

Smit A, Hubeley R, Green P. 1996-2010. RepeatMasker Open-3.0. http://www.repeatmasker.org.

Smith NGC, Webster MT, Ellegren H. 2002. Deterministic mutation rate variation in the human genome. Genome Res. 12(9):1350–1356.

Stark A, et al. 2007. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature. 450(7167):219–232.

Steffen P, Voss B, Rehmsmeier M, Reeder J, Giegerich R. 2006. RNAshapes: an integrated RNA analysis package based on abstract shapes. Bioinformatics. 22(4):500–503.

Stenoien DL, Simeoni S, Sharp ZD, Mancini MA. 2000. Subnuclear dynamics and transcription factor function. J Cell Biochem Suppl. Suppl 35:99–106.

Stephan W. 1996. The rate of compensatory evolution. Genetics. 144(1):419–426.

Stephan W, Kirby DA. 1993. RNA folding in Drosophila shows a distance effect for compensatory fitness interactions. Genetics. 135(1):97–103.

Stoletzki N. 2008. Conflicting selection pressures on synonymous codon use in yeast suggest selection on mRNA secondary structures. BMC Evol Biol. 8:224.

Stoletzki N, Eyre-Walker A. 2007. Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. Mol Biol Evol. 24(2):374–381.

Subramanian AR, Kaufmann M, Morgenstern B. 2008. DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment. Algorithms Mol Biol. 3:6.

Taft RJ, et al. 2009. Tiny RNAs associated with transcription start sites in animals. Nat Genet. 41(5):572–578.

Taft RJ, et al. 2010. Nuclear-localized tiny RNAs are associated with transcription initiation and splice sites in metazoans. Nat Struct Mol Biol. 17(8):1030–1034.

Tellier A, Fischer I, Merino C, Xia H, Camus-Kulandaivelu L, Städler T, Stephan W. 2011. Fitness effects of derived deleterious mutations in four closely related wild tomato species with spatial structure. Heredity. doi: 10.1038/hdy.2010.175.

Thompson JD, Higgins DG, Gibson TJ. 1994. Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22(22):4673–4680.

Tinoco I, et al. 1973. Improved estimation of secondary structure in ribonucleic acids. Nat New Biol. 246(150):40–41.

Tinoco I, Bustamante C. 1999. How RNA folds. J Mol Biol. 293(2):271–281.

Tinoco I, Uhlenbeck OC, Levine MD. 1971. Estimation of secondary structure in ribonucleic acids. Nature. 230(5293):362–367.

Tweedie S, et al. 2009. FlyBase: enhancing Drosophila gene ontology annotations. Nucleic Acids Res. 37(Database issue):D555–D559.

Vicoso B, Charlesworth B. 2009. Effective population size and the faster-X effect: an extended model. Evolution. 63(9):2413–2426.

Wallace IM, O'Sullivan O, Higgins DG, Notredame C. 2006. M-Coffee: combining multiple sequence alignment methods with T-Coffee. Nucleic Acids Res. 34(6):1692–1699.

Washietl S, et al. 2007. Structured RNAs in the encode selected regions of the human genome. Genome Res. 17(6):852–864.

Weinreich DM, Rand DM. 2000. Contrasting patterns of nonneutral evolution in proteins encoded in nuclear and mitochondrial genomes. Genetics. 156(1):385–399.

Weissman DB, Feldman MW, Fisher DS. 2010. The rate of fitness-valley crossing in sexual populations. Genetics. 186(4):1389–1410.

Will S, Reiche K, Hofacker IL, Stadler PF, Backofen R. 2007. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. PLoS Comput Biol. 3(4):e65.

Wood SN. 2006. Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC.

Woolfit M, Bromham L. 2003. Increased rates of sequence evolution in endosymbiotic bacteria and fungi with small effective population sizes. Mol Biol Evol. 20(9):1545–1555.

Woolfit M, Bromham L. 2005. Population size and molecular evolution on islands. Proc Biol Sci. 272(1578):2277–2282.

Yamanaka S, Poksay KS, Arnold KS, Innerarity TL. 1997. A novel translational repressor mRNA is edited extensively in livers containing tumors caused by the transgene expression of the apob mRNA-editing enzyme. Genes Dev. 11(3):321–333.

Zarrinkar PP, Williamson JR. 1994. Kinetic intermediates in RNA folding. Science. 265(5174):918–924.

Zuker M. 1994. Prediction of RNA secondary structure by energy minimization. Methods Mol Biol. 25:267–294.

Acknowledgments

I am deeply indebted to my supervisor Prof. Wolfgang Stephan who made this research possible. Thank you for allowing me to freely pursue my research ideas, for your continuous support and for keeping me on the right track during the work on my projects.

I also want to thank John Parsch and Dirk Metzler as well as Aurélien, Daniel, Pavlos and Stefan for their willingness to help with all my strongly-, moderately-, weakly-, and non-scientific questions. I thank Carlos, Létizia and Francesco who were great office mates and definitely made the office a fun place to work. Of course, I also owe my gratitude to all my fellow graduate students and post-docs from the Drosophila and Tomato groups with whom I have shared my scientific career.

A big 'thank you' goes to Prof. Hideki Innan and the members of his group at the Graduate University for Advanced Studies for being great hosts during my research stay in the Summer of 2010, which also allowed me to discover the natural beauty and culture of Japan.

Finally, I thank my family who believed in me and continuously encouraged and supported me on every step of the way.

Curriculum Vitae

Robert Piskol

Education

01/2008–04/2011	PhD Student, Evolutionary Biology Ludwig-Maximilian-University, Munich, Germany
06/2010–08/2010	Visiting student at the SOKENDAI in the group of Hideki Innan, Hayama, Japan
11/2007	Diploma Degree in Bioinformatics (corresponds to M.S.), *with high distinction*
05/2007–11/2007	Diploma Thesis (subject): Evolution of RNA Structures in Introns
10/2002–11/2007	Student of Bioinformatics at the Ludwig-Maximilian-University and Technical University Munich, Germany

Principal Fields of Study:

Biology	Evolutionary Biology, Population Genetics
Bioinformatics	Machine Learning, Structural Bioinformatics
Informatics	Database Development, Knowledge Discovery in Databases

Publications

Robert Piskol, Wolfgang Stephan (2011)
The role of the effective population size in compensatory evolution
Genome Biol. Evol. (accepted with minor revision)

Robert Piskol, Wolfgang Stephan (2011)
Selective Constraints in Conserved Folded RNAs of Drosophilid and Hominid Genomes
Mol. Biol. Evol. 28(4):1519–1529

Ana Laura Gutierrez-Aguilar, Robert Piskol, Michaela Beitzinger, Jia Yun Zhu, Dagmar Kruspe, Attila Aszodi, Markus Moser, Christoph Englert & Gunter Meister (2010)
The small RNA expression profile of the developing murine urinary and reproductive systems
FEBS Letters 584(21): 4426-4434

Robert Piskol, Wolfgang Stephan (2008)
Analyzing the Evolution of RNA Secondary Structures in Vertebrate Introns Using Kimura's Model of Compensatory Fitness Interactions
Mol. Biol. Evol. 25(11): 2483-2492.

Presentations

R.Piskol (Talk)
What are the Factors that Govern the Evolution of RNA Secondary Structures - Insights from Kimura's Model of Compensatory Evolution

Annual Meeting of the Society for the Study of Evolution (SSE), Moscow ID, U.S.A., 06/2009

R.Piskol, W.Stephan (Poster)
Analyzing the Evolution of RNA Secondary Structures in Vertebrate Introns Using Kimura's Model of Compensatory Fitness Interactions
Annual Meeting of the Society for Molecular Biology and Evolution (SMBE), Iowa City IA, U.S.A., 06/2009

R.Piskol (Talk)
What are the Factors that Govern the Evolution of RNA Secondary Structures - Insights from Kimura's Model of Compensatory Evolution
First VW Status Symposium, Muenster, Germany, 02/2009

R.Piskol (Talk)
Evolution of RNA Structures in Introns
EES Conference '08, Munich, Germany, 10/2008

R.Piskol, W.Stephan (Poster)
Analyzing the Evolution of RNA Secondary Structures in Vertebrate Introns Using Kimura's Model of Compensatory Fitness Interactions
German Conference on Bioinformatics (GCB), Dresden, Germany, 09/2008

Awards and Scholarships

JSPS Summer Program 2010 Scholarship: Visiting student in the group of Prof. Hideki Innan at the Graduate University for Advanced Studies (SOKENDAI), 06/2010 - 08/2010

EES Travel Award; Visit of the 2nd EMBO Workshop on Computational RNA Biology in Cargèse (France), 04/2010

International travel award from the Society for the Study of Evolution, 06/2009

EES Award for distinguished Diploma Thesis, Ludwig-Maximilian-University Munich, 10/2008