Conservation of Neutral Substitution Rate and Substitutional Asymmetries in Mammalian Genes

C.F. Mugal*,[1], J.B.W. Wolf[1], H.H. von Grünberg[2], and H. Ellegren[1]

[1]Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden

[2]Institute of Chemistry, Karl-Franzens University Graz, Graz, Austria

*Corresponding author: E-mail: carina.mugal@ebc.uu.se.

Accepted: 22 December 2009 **Associate editor:** Ross Hardison

Abstract

Local variation in neutral substitution rate across mammalian genomes is governed by several factors, including sequence context variables and structural variables. In addition, the interplay of replication and transcription, known to induce a strand bias in mutation rate, gives rise to variation in substitutional strand asymmetries. Here, we address the conservation of variation in mutation rate and substitutional strand asymmetries using primate- and rodent-specific repeat elements located within the introns of protein-coding genes. We find significant but weak conservation of local mutation rates between human and mouse orthologs. Likewise, substitutional strand asymmetries are conserved between human and mouse, where substitution rate asymmetries show a higher degree of conservation than mutation rate. Moreover, we provide evidence that replication and transcription are correlated to the strength of substitutional asymmetries. The effect of transcription is particularly visible for genes with highly conserved gene expression. In comparison with replication and transcription, mutation rate influences the strength of substitutional asymmetries only marginally.

Key words: neutral substitution rate, substitutional strand asymmetries, transcription-induced mutation, gene expression conservation.

Introduction

Mutation is a fundamental process in evolution and constitutes the raw material for natural selection. However, just as the intensity of selection varies both among populations and genomic regions, so does the incidence by which new mutations occur. There is ample evidence for mutation rate variation among lineages, as well as within genomes, including variation among sites, regions, and chromosomes (Ellegren et al. 2003; Tyekucheva et al. 2008). Knowledge about the determinants of mutation rate variation is of crucial importance to many fields in evolutionary biology, including phylogenetic reconstruction, molecular dating, identification of functional noncoding DNA, and the study of adaptive evolution. Gaining a deeper understanding of the genomic features and molecular processes involved in mutation rate variation will thus be needed to devise more accurate models for molecular evolutionary analyses.

Substitution rate estimates at presumably neutral sites can be used as a proxy for mutation rate. Following this approach, several factors affecting the frequency of mutation have been identified. These include GC content, recombination rate, indel density, the distance to telomeres, exon density, DNA methylation, and chromatin structure (Hardison et al. 2003; Arndt et al. 2005; Prendergast et al. 2007; Tian et al. 2008). However, potential causes of variation are complex and interrelated (Arndt et al. 2005; Tyekucheva et al. 2008). Moreover, the situation is complicated by the fact that both strands of the double-stranded DNA are not always equally affected by mutations, which is referred to as substitutional strand asymmetry (Francino and Ochman 1997; Green et al. 2003). Two processes, DNA replication and transcription, are thought to induce a strand bias in substitution rates. DNA replication is carried out semidiscontinuously where the leading strand is synthesized continuously, whereas the lagging strand is a composite of several Okazaki fragments, approximately 100 kb in length (fig. 1). Because of the discontinuous nature of lagging strand synthesis, the parental strand on which the lagging strand is synthesized (i.e., the leading strand of the previous round of replication) spends significant amount of time in a single-stranded state, which makes it vulnerable to mutagenic reactions, such as hydrolytic deamination, oxidation of guanine, and

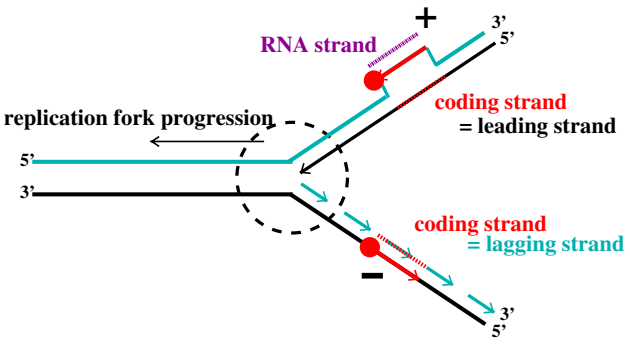

Fig. 1.—Schematic display of the interplay of replication and transcription on transcribed regions of the genome. Both DNA and RNA polymerases (the latter illustrated as a red arrow) synthesize a new DNA or RNA strand in a 5′ to 3′ direction, respectively. They read the DNA strand from 3′ to 5′ direction. Hence, only one daughter strand called leading strand (displayed as a black arrow) is synthesized continuously during replication. The second daughter strand called lagging strand (illustrated by short turquoise arrows) is synthesized discontinuously. The discontinuous lagging strand synthesis leaves its parental DNA strand (a leading strand in the previous round of replication) single-stranded for a significant time and makes it thereby prone to mutations. The coding strand of a gene (illustrated as a dotted red bar) is prone to mutations due to transcription bias. It remains single-stranded for a significant time during active transcription, whereas the noncoding strand hybridizes with the newly synthesized RNA strand. The coding strand can fall within either a leading strand or a lagging strand. If the coding strand is a leading strand, replication and transcription bias are additive, marked with a plus. Replication fork and RNA polymerase progress codirectionally. Replication and transcription bias oppose each other if the coding strand is a lagging strand, marked by a minus. RNA polymerase progresses in the opposite direction of the replication fork.

depurination (Frederico et al. 1990; Grollman and Moriya 1993; Lindahl 1993; Pavlov et al. 2003). Transcription on the other hand may be responsible for strand asymmetric substitution rates in two different ways. First, transcription-coupled repair (TCR) induces a strand bias by preferential repair of bulky lesions on the transcribed, that is, noncoding, strand of the DNA and is thus a rate-reducing mechanism (Mellon 2005; Saxowsky et al. 2008). Second, transcription-induced mutation (TIM) is an acceleration of mutagenic reactions on the coding strand of genes due to its exposure in single-stranded state during active transcription (Beletskii and Bhagwat 1996; Francino and Ochman 2001).

Similar to the neutral substitution rate as such, the degree of substitution rate asymmetries exhibits significant variation across the genome (Mugal et al. 2009). Obviously, if transcription has a strong impact on substitution rate asymmetries, the strand bias should be higher in transcribed regions than in nontranscribed regions, as found, for example, in mammals (Green et al. 2003). Moreover, substitutional asymmetries are more pronounced close to the origin of replication (ORI) in bacterial genomes (Lobry 1996). This is also found for mammalian genomes, where a linear relation between the distance to the nearest ORI and compositional asymmetry has been suggested (Touchon et al. 2005; Mugal et al. 2009). These results may suggest that substitutional asymmetries in transcribed regions are the result of a superposition of transcription- and replication-induced strand asymmetries, as is illustrated in figure 1. However, the relative contribution of each of the two processes is far from being understood (Wang et al. 2008).

Common to several of the factors implicated in governing variation in both neutral substitution rate and substitution rate asymmetries is a certain degree of conservation over time. For example, variation in the local GC content is highly conserved between closely related mammals (Mikkelsen et al. 2005). Moreover, significant correlations between rat–mouse, rat–human, and mouse–human recombination rate have been found (Jensen-Seaman et al. 2004). Furthermore, both processes believed to induce strand asymmetries also exhibit conservation between species: Levels of gene expression are highly conserved between human and mouse (Liao and Zhang 2006; Xing et al. 2007), and the location of the ORIs also shows a high degree of conservation (Cadoret et al. 2008).

Based on the conservation of factors driving or covarying with substitution rate and substitutional asymmetries, it could be expected that substitution rate and substitutional asymmetries themselves are conserved. Surprisingly, little attention has been paid to this question. There is some indication that nucleotide substitution rates are conserved across evolutionary lineages (Smith et al. 2002; Cooper et al. 2004; Mikkelsen et al. 2005; Tyekucheva et al. 2008). In contrast, Imamura et al. (2009) argued that mutation rate preservation between lineages is only weak. The conservation of substitutional asymmetries has to our knowledge not been addressed.

Different approaches based on synonymous sites in genes, transposable elements, or noncoding nonrepetitive sequences can be used to study mutation rate. Although in some cases transposable elements may take on functional roles, they are in general likely to evolve neutrally (Waterston et al. 2002; Ellegren et al. 2003; Hardison et al. 2003; Tyekucheva et al. 2008). Here, we use divergence estimates of transposable elements that have been active after the split of the rodent and primate lineage to assess the degree of conservation in both local mutation rate and substitutional asymmetries between human and mouse orthologs. We estimate lineage-specific mutation rates and substitutional asymmetries of transcribed regions based on repetitive elements, which lie within the intronic regions of the gene. We quantify their degree of conservation and identify the contributing explanatory factors and finally try to disentangle the contribution of mutation rate variation, replication, and transcription on substitutional asymmetries.

Materials and Methods

Statistical Analysis All statistical analyses were performed with the software package R version 2.8.0 (R Development Core Team 2008). Correlations were quantified using Pearson's moment correlation coefficient r. All correlations are significant at a P value threshold of $P < 10^{-4}$ if not explicitly stated otherwise.

Sequence Data Set Human and mouse repeat data were extracted by RepeatMasker version 3.1.2 and RepeatMasker database version 20051025. Genome builds hg18 for human and mm8 for mouse used to extract repeat sequences were downloaded from the University of California—Santa Cruz (UCSC) genome browser (Kent et al. 2002). We excluded low-complexity repeats for the subsequent analysis, which leads to $M = 639$ and $M = 553$ repeat families for human and mouse, respectively, where M denotes the number of repeat families. The RepeatMasker data provide the reconstructed ancestor of each repeat family $\alpha \in \{1, \ldots, M\}$ inserted into the genome at time t_α, as well as a pairwise alignment between each extant repeat copy with its respective ancestral sequence. Primate- and rodent-specific repeat elements were determined by comparing RepeatMasker output of human, mouse, rat, and dog.

Positions of transcribed regions on human and mouse autosomes as well as the assignments of coding strand and noncoding strand were extracted from "KnownGene" at the UCSC table browser (Karolchik et al. 2004). Orthology between human and mouse was established through "hgBlastTab," where the set of genes was restricted to 1:1 orthologs for all subsequent analyses.

Estimation of Local Mutation Rates Local mutation rates for transcribed regions γ were computed using the alignments between intronic repeat copies and their respective ancestral sequences. It is the same method as used in the work of Karro et al. (2008), where the method and its validation are explained in more detail. Our underlying model is a nonhomogeneous Markov chain, similar to that used in other standard approaches of nucleotide sequence evolution. A four-dimensional time-dependent state vector represents the probability of repeat family α being in one of the four states $E = \{A,C,G,T\}$ within region γ. The state vector evolves according to a 4×4 substitution rate matrix q_γ, which contains the 12 independent substitution rates for all possible mutual replacements within the group of the four base nucleotides A, C, G, and T. The transition probability matrix propagating a state vector a time distance $d_{\alpha\gamma}$ forward in time reads

$$P_{\alpha\gamma} = \exp(d_{\alpha\gamma} q_\gamma). \quad (1)$$

Based on the repeat alignments, we first computed the transition probability matrices $P_{\alpha\gamma}$ for all copies of repeat families α located within region γ, where alignment positions involving gaps were discarded. As a measure of sequence divergence of repeat family α in region γ, we then computed the LogDet time distance (Barry and Hartigan 1987),

$$d_{\alpha\gamma} = -\frac{1}{4} \ln\det P_{\alpha\gamma}. \quad (2)$$

One should remark that the substitution rate matrix q_γ is scaled such that its trace is independent of γ and always equal to -4 to ensure consistency between equations (1) and (2) (Karro et al. 2008). Next, we computed family-specific genome-wide averages of the LogDet time distance d_α. We then compared the estimates of divergence $d_{\alpha\gamma}$ of copies of repeat families α located within region γ with their family-specific genome-wide averages d_α. The relative local mutation rate τ_γ was finally defined as the average over the relative differences in divergence weighted by the relative length of the repeat family,

$$\tau_\gamma = \frac{1}{L_\gamma} \sum_\alpha l_{\alpha\gamma} \frac{d_{\alpha\gamma} - d_\alpha}{d_\alpha}. \quad (3)$$

Here, $l_{\alpha\gamma}$ is the total length of all copies of repeat family α in region γ and L_γ is the concatenated length of all repeat copies located in region γ.

Finally, we restricted our analysis to those genes that fulfilled the criterion of containing at least 40 repeat copies within their introns, a threshold set to reduce stochastic variation. Only lineage-specific repeat copies were considered to explore the conservation of mutation rate between human and mouse orthologs. The constraint of lineage specificity was omitted for investigating the relation between local mutation rate and substitutional asymmetries.

Estimation of Substitution Rate Asymmetries We computed the 12 independent substitution rates for transcribed regions γ by estimating the substitution rate matrix q_γ through a maximum-likelihood fit in equation (1). Again, we restricted our analysis to genes containing at least 40 repeat copies within their introns, where lineage specificity was requested to explore the conservation of substitutional asymmetries. As a further constraint, we only considered repeat copies with a minimum length of 50 bp.

To measure the extent of substitutional asymmetries within region γ, we introduce $\omega_{\gamma,X\to Y}$, the relative difference of rate $X \to Y$ on the coding strand and on the noncoding strand,

$$\omega_{\gamma,X\to Y} = \frac{[X\to Y]_{\text{cod}} - [X\to Y]_{\text{noncod}}}{[X\to Y]_{\text{mean}}}. \quad (4)$$

In case the substitution rate $X \to Y$ is higher on the coding strand than on the noncoding strand $\omega_{\gamma,X\to Y} > 0$. $\omega_{\gamma,X\to Y}$

≈ 0 indicates that the rate $X \rightarrow Y$ is almost strand symmetric, whereas negative values point to an opposite trend of substitutional asymmetries, that is, substitution rate $X \rightarrow Y$ is lower on the coding strand than on the noncoding strand.

We focused our analysis on the three most frequent substitutions (connected to its complementary substitutions): the transversion G→T (complementary to C→A) and the two transitions A→G (complementary to T→C) and C→T (complementary to G→A). As an average of rate asymmetry, we computed the arithmetic mean,

$$\omega_\gamma = \frac{1}{3}(\omega_{\gamma,G\rightarrow T} + \omega_{\gamma,A\rightarrow G} + \omega_{\gamma,C\rightarrow T}). \quad (5)$$

Gene Expression Data Set We used Affymetrix exon array expression data of testis for human and mouse genes (Xing et al. 2007) evaluated by Xing et al. (2006) using a probe selection algorithm to determine the level of transcriptional activity in germ cells. We used exon array expression data because it has recently been suggested that such data provide accurate assessments of gene expression allowing comparative studies of gene expression (Xing et al. 2007). Three repeated measurements of human and mouse germ line gene expression values, denoted as expression indices, were available. Following Xing et al. (2007), we took the logarithm of the expression indices to get a measure approximately linearly proportional to transcription levels in the germ cells, denoted as ε. Subsequently, mean values and standard deviations were computed for each set of repeated measurements. Assignments of expression values to known genes were extracted from the UCSC table browser.

Conservation of Gene Expression To assess the degree of conservation in gene expression, we compared human and mouse transcription levels ε. We performed principal component analysis (PCA) to compute the leading PCA line, a line minimizing the residuals of both variables, that is, ε for human and mouse. We then defined δ_ε, a measure of divergence in gene expression as the absolute value of the orthogonal distance of the data point from the leading PCA line.

Replication Bias For any region γ, substitution rates G→T, A→G, and C→T tend to be on average higher on the leading strand than on the lagging strand (Mugal et al. 2009). Thus, ω_γ is on average greater than zero if the differences in substitution rates are calculated with respect to the leading strand and less than zero if the differences are calculated with respect to the lagging strand. Furthermore, the replication-induced strand bias between two adjacent ORIs decreases linearly to zero halfway to the next ORI (Touchon et al. 2005). To assess the influence of replication on substitution rates in transcribed regions, we restricted our analysis to the set of human genes located within 1 of the 678 "N-domains" identified by Huvet et al. (2007), where start and stop positions of an "N-domain" give putative locations of two adjacent ORIs. For each of the genes, we calculated the relative distance $b \in [-1,1]$ between the center of the gene and the center of the N-domain that contains the gene. Note that the center of the N-domain represents the center between two adjacent ORIs, where replication bias is assumed to be zero. We then multiplied b by $+1$ if the coding strand of the gene was placed on the leading strand and by -1 if it was placed on the lagging strand in order to distinguish between replication bias on leading and lagging strands. This weighted distance was denoted as β.

Model Selection Model selection was primarily based on Akaike's information criterion (AIC) (Akaike 1974). Sample sizes were generally large enough ($\frac{n}{K_{global}} > 40$) to use AIC, where n represents the sample size and K_{global} the number of parameters in the model. In order to assess the relative likelihood of competing candidate models, we report normalized Akaike weights *w*AIC. In addition, as AIC has a tendency of overfitting, we used backward selection approaches and the Schwarz information criterion (BIC) that more strongly penalizes the number of parameters (Schwarz 1978). Analogously, we computed normalized weights *w*BIC for model selection based on BIC. In most cases, all three approaches yielded the same results. We explicitly mention where they disagreed.

Results

Conservation of Mutation Rate We analyzed the conservation of relative mutation rate τ_γ across transcribed regions of human and mouse. We computed lineage-specific mutation rates of transcribed regions γ located on human and mouse autosomes by using primate- and rodent-specific repeat elements, respectively. This yielded a set of 197 human and mouse 1:1 orthologous genes, which contained on average 66 primate- and 112 rodent-specific repeat copies within their introns, respectively. Lineage-specific values of τ_γ of human and mouse orthologs were significantly correlated with each other ($r = 0.30$), suggesting local mutation rate conservation between species.

Conservation of Substitution Rate Asymmetries We determined lineage-specific substitution rates of the transcribed regions γ of human and mouse genomes and computed the rate asymmetries for the three most frequent substitutions (and its complementary substitutions): the transversion G→T (complementary to C→A) and the two transitions A→G (complementary to T→C) and C→T (complementary to G→A). In agreement with previous results for mammalian genomes (Green et al. 2003; Mugal et al. 2009),

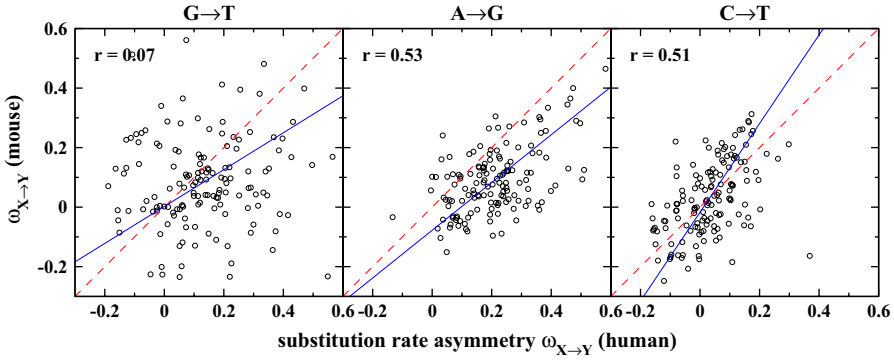

Fig. 2.—Comparison of substitution rate asymmetries $\omega_{X\to Y}$ for the three substitutions G→T, A→G, and C→T between human and mouse orthologs. One dot represents a gene present as a unique copy in each of the two species. The red dashed line represents the bisector line, $y = x$, and the blue solid line represents the leading PCA of $\omega_{X\to Y}$ between human and mouse.

we find that substitution rates of transcribed regions tend to be strand asymmetric. Common to both species, the rates of G→T, A→G, and C→T are on average higher on the coding strand than on the noncoding strand, that is, $\omega_{\gamma,X\to Y}$ is on average greater than zero ($X\to Y$ denotes any of the three substitutions considered). However, asymmetries of individual genes range from −0.23 to 0.58, that is, including genes with negative values of $\omega_{\gamma,X\to Y}$.

To find out if the extent of substitutional asymmetry is conserved between species, we compared rate asymmetries of 142 human and mouse orthologs, which contained on average 64 primate- and 77 rodent-specific repeat copies within their introns, respectively. The correlations of rate asymmetries between species for the three substitutions G→T, A→G, and C→T are shown in figure 2. Comparisons of transitions A→G and C→T reveal a significant positive relationship between substitutional strand asymmetries in human and mouse, with $r = 0.53$ and $r = 0.51$, respectively. This provides evidence that rate asymmetries, and thereby the factors driving them, are conserved between species. No significant interspecies correlation is found for G→T transversion rate asymmetry. Furthermore, comparison of the leading PCA line between the substitution rate asymmetries $\omega_{\gamma,X\to Y}$ in human and mouse, represented as a blue solid line in figure 2, to the bisector line (red dashed line) reveals that substitution rate asymmetries $\omega_{\gamma,G\to T}$ and $\omega_{\gamma,A\to G}$ tend to be stronger in human, whereas $\omega_{\gamma,C\to T}$ shows the opposite trend.

The Effect of Gene Expression Conservation on Substitution Rate Asymmetries Next, we addressed the relationship between substitutional strand asymmetries and male germ line transcription levels for human and mouse. We analyzed a set of 2,554 human genes and 1,253 mouse genes where information of substitution rate asymmetries and transcription levels was available. All three asymmetries $\omega_{G\to T}$, $\omega_{A\to G}$, and $\omega_{C\to T}$ show significant positive correlations with ε = log(expression index) in both species (fig. 3). The strongest correlation is found for the A→G substitution rate asymmetry.

We then asked if the correlation between strand asymmetries and gene expression index is particularly strong

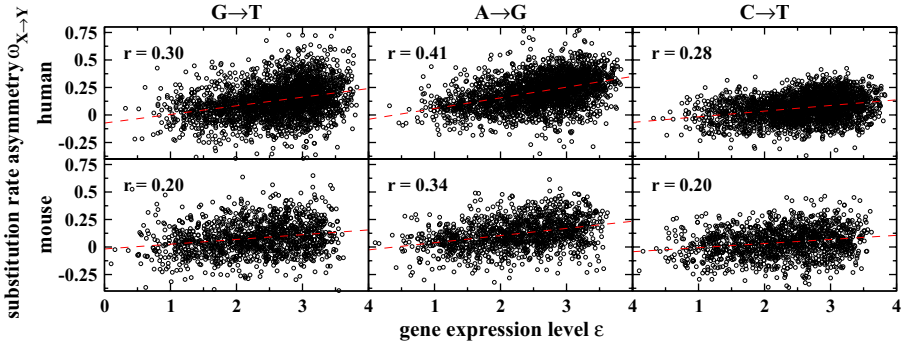

Fig. 3.—Correlation between substitution rate asymmetries and germ line gene expression. The three upper panels show the rate asymmetry for the three substitutions G→T, A→G, and C→T against ε, the logarithm of the gene expression index, for human genes. The three lower panels show the same for mouse genes. One black dot represents one gene. The red dashed line represents the linear regression line. Pearson correlation coefficients between ε and the respective rate asymmetries are shown in each panel.

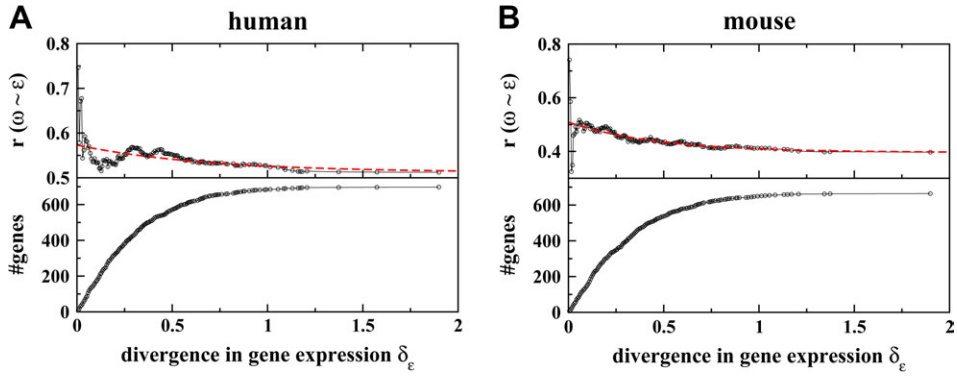

Fig. 4.—Dependence of the correlation between ω and ε [$r(\omega\sim\varepsilon)$] on the degree of conservation of gene expression for an initial set of 698 human genes (4A) and 664 mouse genes (4B). The *x* axis represents the maximum value of gene expression divergence δ_ε between human and mouse in the subset of genes considered. Starting at the very left side, only genes with highly conserved expression levels are depicted. As one moves along this axis to the right, ever more genes with ever more different expression levels between human and mouse are included. The upper panel shows the correlation between gene expression and rate asymmetry as a function of gene expression divergence. One black dot represents the correlation for one subset of genes. The red dashed line is an exponential fit to the data. The lower panel shows the number of genes that are included in the respective subset.

for genes whose expression levels have been conserved over evolutionary time, that is, which show similar levels of expression in human and mouse. Figure 4 indeed suggests that the degree of substitutional rate asymmetry ω does not only scale positively with expression level but also is specifically increased for genes that are conserved in their level of expression. Correlation coefficients *r* reach approximately 0.6 and 0.5 for human and mouse, respectively, according to the exponential fit at maximal similarity in level of gene expression, that is, $\delta_\varepsilon = 0$.

Potential Causes of Substitution Rate Asymmetries In case the coding strand is the leading strand, replication and transcription both work in the same direction and their effect on substitutional asymmetries is expected to be additive. However, when the coding strand is the lagging strand, their effects oppose and may result in reduced observable asymmetries (fig. 1). To dissect the relative contributions of replication and transcription on the extent of substitutional asymmetries, we used a set of 492 human genes where information for all three potentially explanatory variables—the distance of the gene to its nearest ORI (variable β), germ line expression level (variable ε), and the relative mutation rate (variable τ)—is available. For these genes, we calculated the average substitutional rate asymmetry ω. We defined various models with ω as a linear function of different combinations of the three explanatory variables ε, β, and τ and its two-term interactions. The complete list of candidate models and their degree of freedom are provided as supplementary material S1 (Supplementary Material online). We used model selection criteria such as AIC and BIC as well as backward selection to select the most favored models. The four models with the highest explanatory power are

$$Y_1 : \omega = a_0 + a_1\varepsilon + a_2\beta + C \quad (6a)$$

$$Y_2 : \omega = a_0 + a_1\varepsilon + a_2\beta + a_4\varepsilon\beta + C \quad (6b)$$

$$Y_3 : \omega = a_0 + a_1\varepsilon + a_2\beta + a_3\tau + C \quad (6c)$$

$$Y_4 : \omega = a_0 + a_1\varepsilon + a_2\beta + a_3\tau + a_4\varepsilon\beta + a_5\varepsilon\tau + a_6\beta\tau + C, \quad (6d)$$

where $\alpha_0, \ldots, \alpha_6$ represent the model parameters and *C* denotes the error term. In table 1, AIC and BIC values of these four models are listed as well as its differences to its lowest value ΔAIC and ΔBIC, respectively. The relative likelihood of the model using AIC and BIC is represented by its normalized weights *w*AIC and *w*BIC, respectively.

Model selection reveals that both the germ line gene expression level, measured by ε, and the distance to the nearest ORI, denoted as β, have a substantial effect (table 1). The

Table 1
AIC, BIC, and r^2 for the Four Most Favored Models by Model Selection Based on AIC, BIC, and Backward Selection

Model	Degree of Freedom	AIC	ΔAIC	*w*AIC	BIC	ΔBIC	*w*BIC	r^2
Y_1	4	−1,087.32	0.00	0.484	−1,070.39	0.00	0.896	0.39
Y_2	5	−1,086.17	1.15	0.272	−1,065.01	5.38	0.061	0.39
Y_3	5	−1,085.48	1.84	0.192	−1,064.31	6.08	0.043	0.39
Y_4	8	−1,082.86	4.46	0.052	−1,049.00	21.39	0.000	0.39

additive model Y_1 including ε and β is preferred by selection based on AIC and BIC and explains a considerable part of the overall variance in ω ($r^2 = 0.39$). Moreover, considering the whole set of candidate models, it is clear that compared with transcription level and the distance to the nearest ORI, the relative mutation rate τ plays only a subordinate role. Akaike weights (*w*AIC) for ε, β, and τ are 1, 1, and 0.244, respectively. In conclusion, both the transcription level and the distance to the nearest ORI have a substantial impact on substitutional strand asymmetry with comparable strength. Mutation rate alone shows only minor effects. However, it seems to slightly influence the two other explanatory variables, as model Y_4 still is weakly supported (eq. 6).

Discussion

Based on analysis of transposable elements, we computed lineage-specific neutral substitution rates of orthologous genes of human and mouse. We found significant but weak correlations of mutation rates between the two species. Next, we estimated the degree of substitution rate asymmetries in transcribed regions of human and mouse and found that variation in rate asymmetries is more strongly preserved between the two species. The latter finding motivated the subsequent analysis of the causes of variation in substitutional asymmetries, revealing that both transcription and replication have a significant impact on substitutional asymmetries in transcribed regions. Moreover, we provide evidence that the relationship between substitutional asymmetries and transcription depends on the conservation of gene expression. Furthermore, compared with replication and transcription, mutation rate per se has only a marginal influence on the strength of substitutional asymmetries.

Conservation of Substitution Rate and Substitutional Asymmetries Comparison of lineage-specific substitution rates of orthologous regions has provided evidence that variation in mutation rate is conserved between species (Smith et al. 2002; Cooper et al. 2003, 2004; Mikkelsen et al. 2005; Tyekucheva et al. 2008). This has led to the conclusion that substitution rate variation must be deterministic, that is, determined by local genomic features, such as, sequence context (Smith et al. 2002). This hypothesis was further supported by an observed correlation between local mutation rate and GC content (Arndt et al. 2005; Karro et al. 2008). The rare events of regional shifts in the mutation pattern between closely related species, such as mouse and rat or human and chimpanzee, have primarily been related to a bias in GC→AT versus AT→GC mutation rate explained by either a regional mutation bias or a biased gene conversion, leading to dispersing evolution in GC content (Cooper et al. 2004; Ebersberger and Meyer 2005).

However, not only the local sequence context but also factors such as recombination rate, chromatin structure, DNA methylation, and exon density covary with the local rate of mutation (Hardison et al. 2003; Arndt et al. 2005; Prendergast et al. 2007; Tian et al. 2008). These factors are themselves interrelated and also correlated with the local GC content. Thus, unraveling the causal factor of mutation rate variation continues to be a challenge in molecular evolution. Moreover, Hodgkinson et al. (2009) recently found that independently derived single-nucleotide polymorphisms at orthologous sites of human and chimpanzee coincide more often than expected by chance, scaling down conservation of mutation rate variation to the nucleotide level. This finding could not be explained by nearest neighbor effects but rather by more complex sequence context effects, prompting the term "cryptic variation" in mutation rate.

In a recent study based on orthologous repetitive elements, Imamura et al. (2009) questioned the explanatory power of mutation rate conservation. They found a significant but weak correlation of mutation rate between mammalian lineages. In a different approach based on lineage-specific repetitive elements, we here find that local mutation rate is preserved between human and mouse, though explaining less than 10% of the overall variance. These results indicate that local mutation rate might be strongly affected by transient processes, like, recombination hot spots (Jeffreys and Neumann 2009). In contrast, our comparison of substitutional asymmetries between human and mouse orthologs revealed that asymmetries in transitions C→T and A→G show significant conservation over time. Hence, it seems that substitutional asymmetries tend to be less influenced by short-lived processes, as discussed below.

Potential Causes of Substitution Rate Asymmetries During active transcription, the coding strand of genes is exposed in a single-stranded state (Francino et al. 1996; Francino and Ochman 1997; Beletskii and Bhagwat 1998). This potentially results in TIM, which should lead to an acceleration of mutations induced by hydrolytic deamination (C→T and A→G), depurination (A→T and G→T), and oxidation of guanine (G→T) on the coding strand of genes (Beletskii and Bhagwat 1996; Francino and Ochman 2001). An additional transcription-associated mechanism, which is expected to enhance substitutional asymmetries in genes, is TCR. It reduces substitution rates on the noncoding strand by preferential repair of bulky lesions on the noncoding strand but not on the coding strand during active transcription (Oller et al. 1992; Svejstrup 2002; Mellon 2005).

In agreement with the prediction from TIM and TCR, it has recently been shown that on average, the rates of substitutions G→T, A→G, and C→T are increased on the coding strand and decreased on the noncoding strand (Mugal et al. 2009). Our work further supports this finding

by showing that the extent of these substitutional asymmetries is significantly correlated to the level of transcription in the male germ line (fig. 3). This provides evidence that substitutional asymmetries in genes are indeed induced by transcription and are in good agreement with a report of Majewski (2003), who has shown that there is a correlation between the average expression level of housekeeping genes and compositional strand asymmetries. Based on the assumption that the average expression level of housekeeping genes reflects the germ line expression level of this set of genes, he concluded that "compositional" asymmetries are induced by transcription. However, our results suggest a more intricate explanation. Using a direct measure of transcriptional activity in the germ line, we examined the relationship between germ line gene expression and "substitutional" asymmetries. Because substitutional asymmetries can be assumed to be the decisive causal factor behind compositional asymmetries (Mugal et al. 2009), our results suggest that differing expression levels lead via TIM and TCR to substitutional asymmetries of different strengths and subsequently entail compositional asymmetries. Thus, the correlation between substitutional asymmetries and transcriptional activity in the germ line provides the logical link for the correlation found by Majewski. Furthermore, by comparing the strength of correlations between germ line gene expression and substitutional asymmetries $\omega_{G\to T}$, $\omega_{A\to G}$, and $\omega_{C\to T}$, we find evidence that transcription most strongly biases the A→G substitution rate.

Nevertheless, one should bear in mind that the level of transcription merely represents the present activity of gene expression, whereas the estimated degree of substitutional asymmetries is a time-averaged quantity, averaged over the period between the age of the youngest and the age of the oldest repeat family used in the estimation procedure. If gene expression levels strongly varied over this time period, we should not expect a strong correlation. However, previous findings suggest that expression levels have remained fairly constant since the split of human and mouse (Liao and Zhang 2006; Xing et al. 2007). Those genes having a highly conserved gene expression level can be expected to have less variable substitutional asymmetries over time. Hence, time-averaged substitutional asymmetries should show a better correlation to germ line transcription level. This is indeed shown in figure 4: The higher the degree of gene expression conservation, the higher the strength of correlation. This again supports the hypothesis that substitution rate asymmetries in genes are induced by transcription. Furthermore, it shows that short-lived processes, that is, unstable gene expression, have weaker influence on the average substitution rate asymmetries than conserved processes.

Transcription-induced strand asymmetries can only be found in those regions of the genome that are actively transcribed in the germ line. We here focus our analysis on protein-coding genes. However, recent development of techniques such as RNA deep sequencing and chromatin immunoprecipitation shows that transcription is more pervasive than previously expected (Jacquier 2009). Hence, transcription affects larger parts of the genome, and TIM bias may therefore not be restricted to protein-coding regions only. Moreover, replication affects substitution rates in the whole genome, where the strongest strand bias is found close to the ORIs (Touchon et al. 2005). During lagging strand synthesis, the leading strand remains single stranded for a significant time, which makes it vulnerable to reactions, such as hydrolytic deamination, depurination, and oxidation of guanine (Frederico et al. 1990; Grollman and Moriya 1993; Lindahl 1993; Pavlov et al. 2003). As a consequence in 3′ direction of an ORI, that is, on the leading strand, the DNA strand is enriched in guanine (G) and thymine (T), whereas in 5′ direction, that is, on the lagging strand, the strand is depleted in G and T. In bacteria, this strand bias in nucleotide composition is so pronounced that compositional skew diagrams are often deployed to identify the locations of ORIs (Grigoriev 1998). Measuring the influence of replication on substitution rate asymmetries in mammalian genomes is complicated by the fact that 1) there exist several ORIs 2) that may not all be used during one cell division and 3) whose exact locations are often not known. Nevertheless, variation in substitutional asymmetries should be related to the distance of the region under study to its nearest ORI and in case of transcribed regions to its level of germ line gene expression. Based on results of cumulative skew diagram analysis, Wang et al. (2008) suggested that replication plays only a minor role in large vertebrate genes. Here, we used a different approach based on recent computational advances that allowed for the identification of some putative ORIs in the human genome (Huvet et al. 2007). We analyzed the relationship between substitutional asymmetries in genes and the distance of the gene to its nearest ORI as well as its germ line transcription level. The results, summarized in table 1, provide evidence that both processes have substantial impact on substitutional asymmetries.

Interestingly, when comparing replication- and transcription-induced strand asymmetries of substitution rates $X\to Y$, the strongest correlation is found for the C→T substitution rate asymmetry and the distance to the nearest ORI (Mugal et al. 2009), whereas the strongest correlation between substitution rate asymmetry and germ line gene expression level is found for the A→G substitution rate. Both processes, replication and transcription, that induce substitutional strand bias show a high degree of conservation between species (Liao and Zhang 2006; Xing et al. 2007; Cadoret et al. 2008). This is in good agreement with the finding that substitutional asymmetries are conserved between human and mouse orthologs (fig. 2). Our analysis suggests that mutation rate has only a marginal influence on the strength of substitutional asymmetries. However, because preservation

in mutation rate is rather weak, we might not be able to detect the real strength of its effect on substitutional asymmetries.

In conclusion, substitution rates in transcribed regions are significantly affected by replication and transcription, leading to strand asymmetric substitution rates. These asymmetries are thus the result of neutral processes, which has important implications for several aspects of molecular evolution. To mention but one of the most prominent, it may in a similar way also affect 4-fold degenerate codon positions (Qu et al. 2006). It seems relevant to include the effects of substitutional asymmetries in estimates of codon usage bias. Especially, as selection on 4-fold degenerate sites is predominantly found in highly expressed genes that according to the present analysis are most strongly affected by TIM bias.

Supplementary Material

Supplementary material S1 is available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

We are grateful to John Karro for providing software and for his constant support and advice. Thanks to Judith Mank for helpful discussions about gene expression analysis. Special thanks to Ellen Zechner and the Graduate School "Molecular Enzymology" of the University of Graz who supported the research collaboration between the University of Graz and Uppsala University. C.F.M. received financial support from the Graduate School Molecular Enzymology of the University of Graz. H.E. is supported by the Swedish Research Council and the Knut and Alice Wallenberg Foundation.

Literature Cited

Akaike H. 1974. New look at statistical-model identification. Ieee Trans Automat Contr. 19:716–723.

Arndt PF, Hwa T, Petrov DA. 2005. Substantial regional variation in substitution rates in the human genome: importance of GC content, gene density, and telomere-specific effects. J Mol Evol. 60:748–763.

Barry D, Hartigan JA. 1987. Asynchronous distance between homologous DNA-sequences. Biometrics. 43:261–276.

Beletskii A, Bhagwat AS. 1996. Transcription-induced mutations: increase in C to T mutations in the nontranscribed strand during transcription in Escherichia coli. Proc Natl Acad Sci U S A. 93: 13919–13924.

Beletskii A, Bhagwat AS. 1998. Correlation between transcription and C to T mutations in the non-transcribed DNA strand. Biol Chem. 379: 549–551.

Cadoret J-C, et al. 2008. Genome-wide studies highlight indirect links between human replication origins and gene regulation. Proc Natl Acad Sci U S A. 105:15837–15842.

Cooper GM, et al. 2003. Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. Genome Res. 13:813–820.

Cooper GM, et al. 2004. Characterization of evolutionary rates and constraints in three mammalian genomes. Genome Res. 14: 539–548.

Ebersberger I, Meyer M. 2005. A genomic region evolving toward different GC contents in humans and chimpanzees indicates a recent and regionally limited shift in the mutation pattern. Mol Biol Evol. 22:1240–1245.

Ellegren H, Smith NGC, Webster MT. 2003. Mutation rate variation in the mammalian genome. Curr Opin Genet Dev. 13:562–568.

Francino MP, Chao L, Riley MA, Ochman H. 1996. Asymmetries generated by transcription-coupled repair in enterobacterial genes. Science. 272:107–109.

Francino MP, Ochman H. 1997. Strand asymmetries in DNA evolution. Trends Genet. 13:240–245.

Francino MP, Ochman H. 2001. Deamination as the basis of strand-asymmetric evolution in transcribed Escherichia coli sequences. Mol Biol Evol. 18:1147–1150.

Frederico LA, Kunkel TA, Shaw BR. 1990. A sensitive genetic assay for the detection of cytosine deamination—determination of rate constants and the activation-energy. Science. 29:2532–2537.

Green P, et al. 2003. Transcription-associated mutational asymmetry in mammalian evolution. Nat Genet. 33:514–517.

Grigoriev A. 1998. Analyzing genomes with cumulative skew diagrams. Nucleic Acids Res. 26:2286–2290.

Grollman AP, Moriya M. 1993. Mutagenesis by 8-oxoguanine: an enemy within. Trends Genet. 9:246–249.

Hardison RC, et al. 2003. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. Genome Res. 13:13–26.

Hodgkinson A, Ladoukakis E, Eyre-Walker A. 2009. Cryptic variation in the human mutation rate. PLoS Biol. 7:e27EP.

Huvet M, et al. 2007. Human gene organization driven by the coordination of replication and transcription. Genome Res. 17:1278–1285.

Imamura H, Karro J, Chuang J. 2009. Weak preservation of local neutral substitution rates across mammalian genomes. BMC Evol Biol. 9:89.

Jacquier A. 2009. The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. Nat Rev Genet. 10: 833–844.

Jeffreys AJ, Neumann R. 2009. The rise and fall of a human recombination hot spot. Nat Genet. 41:625–629.

Jensen-Seaman MI, et al. 2004. Comparative recombination rates in the rat, mouse, and human genomes. Genome Res. 14:528–538.

Karolchik D, et al. 2004. The UCSC table browser data retrieval tool. Nucleic Acids Res. 32:D493–D496.

Karro JE, Peifer M, Hardison RC, Kollmann M, von Grünberg HH. 2008. Exponential decay of GC content detected by strand-symmetric substitution rates influences the evolution of isochore structure. Mol Biol Evol. 25:362–374.

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

Liao B-Y, Zhang J. 2006. Evolutionary conservation of expression profiles between human and mouse orthologous genes. Mol Biol Evol. 23:530–540.

Lindahl T. 1993. Instability and decay of the primary structure of DNA. Nature. 362:709–715.

Lobry JR. 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol. 13:660–665.

Majewski J. 2003. Dependence of mutational asymmetry on gene-expression levels in the human genome. Am J Hum Genet. 73: 688–692.

Mellon I. 2005. Transcription-coupled repair: a complex affair. Mutat Res-Fund Mol M. 577:155–161.

Mikkelsen TS, et al. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature. 437:69–87.

Mugal CF, von Grünberg H-H, Peifer M. 2009. Transcription-induced mutational strand bias and its effect on substitution rates in human genes. Mol Biol Evol. 26:131–142.

Oller AR, Fijalkowska IJ, Dunn RL, Schaaper RM. 1992. Transcription-repair coupling determines the strandedness of ultraviolet mutagenesis in Escherichia coli. Proc Natl Acad Sci U S A. 89: 11036–11040.

Pavlov YI, Mian IM, Kunkel TA. 2003. Evidence for preferential mismatch repair of lagging strand DNA replication errors in yeast. Curr Biol. 13:744–748.

Prendergast J, et al. 2007. Chromatin structure and evolution in the human genome. BMC Evol Biol. 7:72.

Qu H-Q, Lawrence S, Guo F, Majewski J, Polychronakos C. 2006. Strand bias in complementary single-nucleotide polymorphisms of transcribed human sequences: evidence for functional effects of synonymous polymorphisms. BMC Genomics. 7:213.

R Development Core Team. 2008. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Saxowsky TT, Meadows KL, Klungland A, Doetsch PW. 2008. 8-Oxoguanine-mediated transcriptional mutagenesis causes Ras activation in mammalian cells. Proc Natl Acad Sci U S A. 105:18877–18882.

Schwarz G. 1978. Estimating dimension of a model. Ann Stat. 6:461–464.

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

Svejstrup JQ. 2002. Mechanisms of transcription-coupled DNA repair. Nat Rev Mol Cell Bio. 3:21–29.

Tian D, et al. 2008. Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. Nature. 455:105–108.

Touchon M, et al. 2005. Replication-associated strand asymmetries in mammalian genomes: toward detection of replication origins. Proc Natl Acad Sci U S A. 102:9836–9841.

Tyekucheva S, et al. 2008. Human-macaque comparisons illuminate variation in neutral substitution rates. Genome Biol. 9:R76.

Wang H-F, Hou W-R, Niu D-K. 2008. Strand compositional asymmetries in vertebrate large genes. Mol Biol Rep. 35:163–169.

Waterston RH, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature. 420:520–562.

Xing Y, Kapur K, Wong WH. 2006. Probe selection and expression index computation of affymetrix exon arrays. PLoS ONE. 1:e88.

Xing Y, Ouyang Z, Kapur K, Scott MP, Wong WH. 2007. Assessing the conservation of mammalian gene expression using high-density exon arrays. Mol Biol Evol. 24:1283–1285.