Genomic Regionality in Rates of Evolution Is Not Explained by Clustering of Genes of Comparable Expression Profile

Martin J. Lercher, Jean-Vincent Chamary, and Laurence D. Hurst[1]

Department of Biology and Biochemistry, University of Bath, Bath, BA2 7AY, United Kingdom

In mammalian genomes, linked genes show similar rates of evolution, both at fourfold degenerate synonymous sites (K_4) and at nonsynonymous sites (K_A). Although it has been suggested that the local similarity in the synonymous substitution rate is an artifact caused by the inclusion of disparately evolving gene pairs, we demonstrate here that this is not the case: after removal of disparately evolving genes, both (1) linked genes and (2) introns from the same gene have more similar silent substitution rates than expected by chance. What causes the local similarity in both synonymous and nonsynonymous substitution rates? One class of hypotheses argues that both may be related to the observed clustering of genes of comparable expression profile. We investigate these hypotheses using substitution rates from both human–mouse and mouse–rat comparisons, and employing three different methods to assay expression parameters. Although we confirm a negative correlation of expression breadth with both K_4 and K_A, we find no evidence that clustering of similarly expressed genes explains the clustering of genes of comparable substitution rates. If gene expression is not responsible, what about other causes? At least in the human–mouse comparison, the local similarity in K_A can be explained by the covariation of K_A and K_4. As regards K_4, our results appear consistent with the notion that local similarity is due to processes associated with meiotic recombination.

[Supplemental material is available online at www.genome.org.]

In mammals, it is claimed that the rates of both protein sequence evolution (Williams and Hurst 2000; Lercher et al. 2001) and synonymous nucleotide change (Casane et al. 1997; Matassi et al. 1999; Nachman and Crowell 2000; Lercher et al. 2001; Smith et al. 2002; Yi et al. 2002; Hardison et al. 2003) show local clustering, with neighboring regions evolving at similar rates. However, other authors claim that these results, at least as regards synonymous nucleotide changes, are nothing more than methodological artifacts (Kumar and Subramanian 2002). Here we ask two questions. First, is the local similarity in synonymous substitution rates real? We show that it is. Given this, we then ask why linked genes might have similar synonymous and nonsynonymous substitution rates. In particular, we examine the hypothesis that transcriptional activity provides a possible mechanistic basis for both clustering phenomena (Hurst and Eyre-Walker 2000; Williams and Hurst 2002; Hardison et al. 2003). A priori, a coupling with transcriptional activity is an attractive hypothesis, as genes with comparable expression profile cluster (Caron et al. 2001; Lercher et al. 2002b; Lercher et al. 2003; Versteeg et al. 2003) and expression parameters are related to substitution rates (Duret and Mouchiroud 2000). We show here that transcriptional activity appears not to be an important variable underpinning local similarity of rates of evolution. Finally, we briefly ask what else might then explain the clustering. As the extent of local similarity appears different in the mouse–rat comparison and in the human–mouse comparison (Lercher et al. 2001), we analyze both.

Is Local Similarity an Artifact?

Kumar and Subramanian (2002) argue that all previous findings of local similarity in synonymous substitution rates are invalid, owing to a methodological problem in how one estimates substitution rates. Classical methods (e.g., Tamura and Nei 1993) for estimating substitution rates from pairwise alignments assume that the process of molecular evolution is the same in two sequences since the time of common ancestry. If this assumption is not upheld, the methods will provide biased estimates. Movement of gene clusters into a new environment (e.g., by chromosome rearrangements) can result in the nucleotide composition of the genes "ameliorating" to their new location (Kumar and Subramanian 2002). The inclusion of such sequences with disparate substitution patterns would bias results, leading to overestimates of the actual point substitution rates in the translocated gene cluster (Kumar and Subramanian 2002). Kumar and Gadagkar (2001) developed a disparity test to diagnose such heterogeneity, and suggested that only those alignments passing the test should be employed to investigate mutation rates. Significantly, 46% of human–mouse alignments fail the disparity test, whereas only 8% of mouse–rat alignments are deviant (Kumar and Subramanian 2002).This alone might then explain the observation (Lercher et al. 2001) that local similarity in the synonymous substitution rate is weaker in the mouse–rat than in the human–rodent comparison.

By examining the relationship between the difference in the synonymous substitution rate between two genes and the physical distance between the genes, Kumar and Subramanian (2002) then argued that, in their purified data set, there was no evidence that rates of evolution varied across the genome. Further, the authors argued that all between-gene variation in evolutionary rate is attributable to estimation error owing to differences in length of sequence, and hence that one cannot reject the notion that there is one mutation rate for all autosomal sequences. These and other conclusions from their study have been challenged (e.g., that the global clock rate of synonymous evolution does not differ between mammalian lineages, Yi et al. 2002).

Below, we first confirm that the signals of local similarity

[1]Corresponding author.
E-MAIL l.d.hurst@bath.ac.uk; FAX 44 (0)1225 826779.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.1597404.

14:1002–1013 ©2004 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/04; www.genome.org

examined here are not due to disparate substitution patterns. Additionally, to be more confident that we are examining regionality in nonselective substitution processes, we ask whether, in the mouse–rat analysis, introns from a given gene have more similar rates of evolution than expected by chance. Although thereafter we report results for K_4 only from the subgroup of genes that passed the disparity test, results including all genes are very similar (data not shown). To further minimize the potential effect of disparate evolutionary patterns, we calculate K_4 employing a recently developed scheme that attempts to control for disparity (Tamura and Kumar 2002).

Are Clusters of Germline-Expressed Genes Responsible for Regionality in the Mutation Rate?

The rate of synonymous nucleotide change, assayed at fourfold degenerate sites (K_4), is often assumed to reflect the local mutation rate (but see Eyre-Walker 1999; Smith and Eyre-Walker 2001; Duret et al. 2002; Lercher and Hurst 2002a; Lercher et al. 2002a). Whether more highly expressed genes should be faster or slower evolving at synonymous sites is, however, hard to predict on a priori grounds, as two putative forces have potentially opposite effects. Although there is, for example, evidence that transcription can be mutagenic (Datta and Jinks-Robertson 1995; Aguilera 2002), there are repair mechanisms that are coupled to transcription (Mellon et al. 1987; Selby and Sancar 1993; van Gool et al. 1997). The latter is believed to explain a recently described transcription-associated strand mutational asymmetry in mammals, which has acted to produce a compositional asymmetry, an excess of G+T over A+C on the coding strand, in most genes (Green et al. 2003).

If the two forces (mutation and repair) do not cancel out, a covariation between genic mutation rates and rates of transcription in the germline is to be expected. Prior evidence suggests the possibility of a weak reduction in the synonymous rate of substitution in putatively germline-expressed genes (Duret and Mouchiroud 2000). Given too that it has been shown that highly expressed genes are clustered in the human genome (Caron et al. 2001; Versteeg et al. 2003), transcriptionally mediated variations in the mutation rate could potentially lead to the observation of local similarity.

Are Clusters of Housekeeping Genes Responsible for Regionality in the Rate of Protein Sequence Evolution?

The clustering of highly expressed genes has been shown to be a secondary effect caused by clustering of housekeeping genes (Lercher et al. 2002b). Broadly expressed genes (i.e., those expressed in many tissues, not necessarily at a high rate) are known to evolve at lower rates (Hastings 1996; Duret and Mouchiroud 2000; Williams and Hurst 2002), possibly owing to stronger purifying selection on proteins that have to function in a wide range of different tissues. Thus, such clustering according to breadth of expression might also explain the local similarity in the rate of protein sequence evolution (assayed as K_A, the rate of nonsynonymous substitutions; Williams and Hurst 2002).

Alternative Hypotheses

As regards the above issues, we show that local similarity in both synonymous and nonsynonymous substitution rates is real, but is not explained by the clustering of transcriptionally comparable genes. What other explanations might there be? With regard to the nonsynonymous substitution rate, we ask whether the local similarity is driven by a corresponding local similarity in the synonymous substitution rate. Early claims from the mouse–rat analysis suggested this was not the case (Williams and Hurst 2000), but more recent reanalysis argues to the contrary (Malcom et al. 2003). We return to this issue and ask, why, if local similarity in the nonsynonymous substitution rate is largely owing to underlying variation in the mutation rate, is the effect more pronounced in the vicinity of tissue-specific genes (Williams and Hurst 2002).

Aside from transcription, other possibilities to explain the local variation in the synonymous substitution rate include (1) heterogeneity in the activity of repair enzymes (Matassi et al. 1999), (2) recombination-associated mutational and/or repair hotspots (Perry and Ashworth 1999; Lercher and Hurst 2002b; Filatov and Gerrard 2003; Hellmann et al. 2003), and (3) GC-associated mutation or fixation biases (Lercher et al. 2001; Castresana 2002b; Smith et al. 2002; Yi et al. 2002; Hardison et al. 2003). We examine the last two of these together, as it has been argued that they are not independent (Meunier and Duret 2004).

Methological Issues

Unfortunately, investigations of this nature can suffer a number of problems. First, there is no unambiguously best way to estimate expression parameters (Huminiecki et al. 2003). To have more confidence in any claim that we might wish to make, we use all possible sources of high-throughput data (EST, microarray, and SAGE) so as to test for the robustness of all results. As EST data provide poor representation of expression rates, the latter are estimated from SAGE and microarray data alone. Further, recent evidence suggests that in GC-rich regions, the substitution process may well be affected by both mutation and biased gene conversion (Eyre-Walker 1999; Smith and Eyre-Walker 2001; Duret et al. 2002; Lercher and Hurst 2002a; Lercher et al. 2002a). To distinguish mutation from fixation biases, we also analyze separately GC-poor sequences. Finally, as gene duplications often occur in tandem and as it is possible that the two resulting proteins are under similar purifying selection, neighboring duplicates could also contribute to local similarity in the rate of protein sequence evolution. Although some prior analyses control for the presence of tandem gene duplications (Williams and Hurst 2000, 2002; Lercher et al. 2001), the extent of their potential contribution to local similarity has yet to be evaluated. We report that the effect is substantial even for very distantly related genes, a result which underlines the necessity to eliminate them prior to analysis.

RESULTS AND DISCUSSION

Disparity Is Not Responsible for Local Similarity

Before analyzing any putative causes of local similarity, it is first necessary to establish that such local similarity exists. To test for local similarity, we used a modified version of the method of Lercher et al. (2001). For each gene, we calculated a 'focal average' of the substitution rate, that is, an average over the rates of all other genes within 1 Mb of the focal gene. We denote the correlation coefficient for all data pairs consisting of (1) the rate of the focal gene and (2) the corresponding focal average as the 'focal average correlation', ρ. Thus, the square of ρ estimates what fraction of the variation in the rate K can be explained by comparison to an independent regional average. We estimated statistical significance by a randomization procedure (see Methods). This method was applied to rate estimates of nonsynonymous (K_A) and fourfold degenerate synonymous (K_4) substitutions from orthologous human–mouse and mouse–rat coding sequences, as well as to estimates of intronic point substitution (K_i) and indel (K_{indel}) rates from orthologous mouse–rat introns. Throughout the manuscript, we restrict our analysis of synony-

mous sites to fourfold degenerate third sites; this makes the counting of such sites unambiguous.

Local similarity measured by ρ was significant for synonymous (K_4) and nonsynonymous (K_A) rates, as well as for the intronic substitution (K_i) and indel (K_{indel}) rates (Table 1). Kumar and Subramanian (2002) argued that local similarities in point substitutions may be a methodological artifact, and suggested that genes failing a disparity test should be excluded (Kumar and Gadagkar 2001). We therefore repeated the analysis, this time excluding genes that fail the disparity test ($P_{disparity} < 0.05$, estimated for fourfold degenerate sites and intronic sites, respectively). In approximate agreement with Kumar and Subramanian (2002), we found that 61% of orthologs failed the disparity test in the human–mouse comparison, whereas the corresponding figure for mouse–rat was only 13%. Excluding these disparate genes, we again found significant local similarity for all substitution measures (Table 1, Fig. 1). From the modest reduction in ρ values, we conclude that only a relatively small part of local similarity is caused by the inclusion of disparate genes. To be conservative, all analyzes of K_4 reported below include only nondisparately evolving genes. Very similar results are obtained when including all genes (data not shown). The notion that disparity is not a major cause of local similarity is supported by our finding of a similarly strong local similarity of indel rates, because indels are not expected to suffer from the same estimation problems as point substitutions. Our finding of comparable local similarities for nucleotide substitution and indel rates is consistent with the observation that indels and point substitutions cluster in the same regions (Hardison et al. 2003). Although this suggests that the processes of substitution and insertion/deletion may be mechanistically coupled (Ogata et al. 1996; Hardison et al. 2003), we observe no correlation between the two rates in introns on a within-gene scale (data not shown, c.f. Ogata et al. 1996).

In agreement with the above results, we also found a significant correlation between intronic substitution rates (K_i) and the synonymous substitution rate (K_4) in flanking exons ($r_{Spearman} = 0.17$, $P = 0.030$, both for all genes and for nondisparate genes). The same is reported in the human–mouse comparison after exclusion of fast-evolving sequence (Castresana 2002a). A prior study failed to detect such an effect in the mouse–rat comparison (Hughes and Yeager 1997). It has been suggested that this may be due to a limited sample size (Castresana 2002a), which we supported by a simulated sample size reduction of our data (not shown).

Why does our result differ from that of Kumar and Subramanian (2002), who concluded that controlling for disparity does destroy the signal of local similarity? We believe that the crucial difference lies in the measure of local similarity. Kumar and Subramanian examined the correlation between chromosomal distance and the difference in K_4 across individual, directly neighboring gene pairs. This method has at least two weaknesses. First, it supposes that all of the variation between genes occurs within a chromosomal region and not between chromosomes. If a large part of between-gene variation is actually between chromosomes (Lercher et al. 2001; Castresana 2002b; Ebersberger et al. 2002; Malcom et al. 2003), this method might fail to find any signal. However, the exclusion of between-chromosome effects (by permuting our intronic rates only within the same chromosome) hardly decreased the significance of the local similarity in K_i (data not shown). This confirms the previous notion that a substantial part of local similarity in point substitution rates is independent of chromosomal effects (Lercher et al. 2001).

Further, the method of Kumar and Subramanian may not be sensitive enough for a weak similarity signal. Notably, they considered only individual neighboring gene pairs. As variance in K_4 estimates is dominated by size-dependent noise (Kumar and Subramanian 2002), the value of the difference between two genes will have a large error component; this component may be substantially reduced by the calculation of focal averages. Further, restriction on next neighbors results in very low sample sizes, especially at larger distances. We tested this through an implementation of the method of Kumar and Subramanian (2002). From the human–mouse data set, we first excluded disparately evolving genes. We then calculated the absolute difference in K_4 ($|\Delta K_4|$) between neighboring genes that reside in the same syntenic region. Using a sample size similar to that of Kumar and Subramanian (2002), we confirmed that local similarity is not detectable when comparing mean K_4 across windows of different gene distances (Fig. 1; window size 200 or 500 kb). However, for distances exceeding 1.5 Mb, windows contained inadequate sample sizes well below 50 genes. We suggest therefore that the protocol employed by Kumar and Subramanian (2002) may be too prone to size-dependent error variance to detect what

Table 1. Genomic Estimates of Local Similarity in Evolutionary Rates, Assayed by the Focal Average Correlation ρ

	Human–mouse			Mouse–rat		
	ρ[a]	p[b]	N	ρ[a]	p[b]	N
All genes						
K_A	0.126	<0.0001	4596	0.055	0.0018	4116
K_4	0.334	<0.0001	4284	0.183	<0.0001	3909
K_4^{excCpG}	0.251	<0.0001	4178	0.179	<0.0001	3815
K_i	—	—	—	0.170	<0.0001	541
K_{indel}	—	—	—	0.191	<0.0001	541
Nondisparate genes						
K_A	0.103	0.0003	1568	0.065	0.0010	3554
K_4	0.272	<0.0001	1377	0.166	<0.0001	3380
K_4^{excCpG}	0.189	<0.0001	1292	0.173	0.0002	3298
K_i	—	—	—	0.151	0.0016	469
K_{indel}	—	—	—	0.160	0.0004	469

[a]Focal average correlation = correlation coefficient for data pairs, each consisting of a gene's rate K and the mean of its neighbors within 1 Mb.
[b]Number of equal or higher ρ values found in 10,000 randomly rearranged genomes.

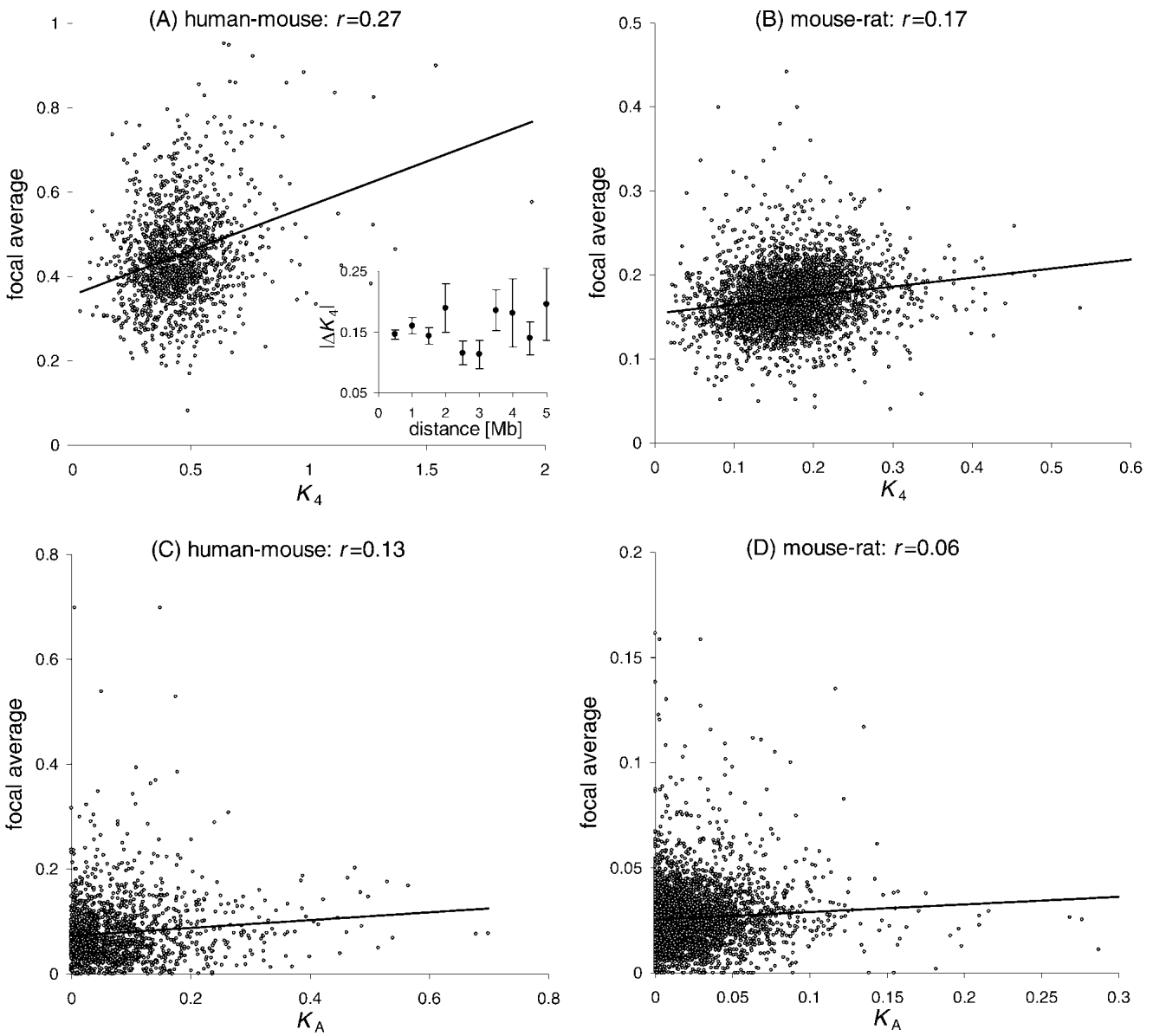

Figure 1 Correlation ρ between substitution rates and focal averages. For K_4 (*A*,*B*), only nondisparate gene pairs are included. All correlations are highly significant ($P \leq 0.001$). The *inset* in (*A*) shows an alternative measure of local similarity as used by Kumar and Subramanian (2002).

are relatively weak effects, and also failed to take account of between-chromosome variation.

Transcription Affects K_4 but Does Not Explain Local Similarity

Any putative mutational processes associated with transcription are only relevant if occurring in the germline, as only these mutations will reach the next generation. Thus, if transcription does affect the mutation rate, then genes transcribed in the germline should show different substitution rates compared to genes not transcribed in the germline. Unfortunately, we do not have large-scale expression data for mammalian germlines. Although genomic expression data does exist for ovaries and testes (Su et al. 2002), these tissues contain a large number of somatic cells. However, we can gain insight into this issue from analysis of putative housekeeping genes (for definitions see Methods), which should also be expressed in germline cells. In contrast, putative tissue-specific genes (expressed in 0 or 1 of our tissues) are unlikely to be expressed in the germline. As predicted from the hypothesis of transcriptionally induced repair, we found that putative housekeeping genes have significantly lower substitution rates at fourfold degenerate sites (Table 2). However, differences are small, suggesting that any transcriptional effect explains only a small part of the observed variation in substitution rates. Consistent with this difference between tissue-specific and housekeeping genes, and in agreement with previous findings (Duret and Mouchiroud 2000), we also found a significant negative correlation between expression breadth and K_4 (Table 3).

Although the latter effect is weak, we can still ask whether the local similarity in K_4 is owing to the clustering of genes of comparable expression breadth (either because more broadly expressed genes are more likely to be expressed also in germline, or for other unknown reasons). To test this hypothesis, we calculated ρ values for K_4 across all genes with valid expression esti-

Table 2. Difference of Synonymous Substitution Rate (K_4) Between Housekeeping and Tissue-Specific Genes

	Human–mouse					Mouse–rat				
Breadth measure	House-keeping[a]	Tissue-specific[a]	ΔK_4[b]	P[c]	N	House-keeping[a]	Tissue-specific[a]	ΔK_4[b]	P[c]	N
K_4										
EST	0.387	0.502	0.115	0.00077	121	0.161	0.18	0.019	0.00067	463
SAGE	0.411	0.475	0.064	0.0026	394	0.157	0.175	0.018	<0.00001	1928
Microarray	0.387	0.501	0.114	0.00005	103	0.165	0.172	0.007	0.057	789
K_4 residuals from regression on K_A										
EST	0.024	0.055	0.047	121	−0.031	0.0099	0.0012	0.41	463	0.0087
SAGE	−0.002	0	0.51	394	−0.002	0.0092	0.0065	0.032	1928	0.0027
Microarray	0.026	0.048	0.05	103	−0.022	0.0067	−0.0012	0.61	789	0.0079

[a]Average K_4 or average residuals of K_4. Residuals were calculated from expected K_4 values, which were predicted from linear regression of log(K_4) on log(K_A) including all genes. Genes were classified as housekeeping/tissue-specific if supported by experiments in both human and mouse for the human–mouse comparison, and by experiments in mouse for the mouse–rat comparison. Only non-disparate gene pairs were included.
[b]Difference in K_4 (or residuals) between tissue-specific and housekeeping averages.
[c]Probability of finding an equal or greater difference in 100,000 randomized genomes.

mates. Statistical significance was estimated by comparison of ρ to 10,000 data sets obtained by randomly permuting the positions of all genes (P_{all}, Table 4). We then repeated the randomization procedure, this time permuting only gene positions within classes of similar expression breadth (P_{group}, Table 4). If local similarity was largely independent of expression breadth, we expect $P_{group} \approx P_{all}$; conversely, if expression breadth determined a large part of the regional variation in K_4, we expect $P_{group} \gg P_{all}$, as randomization in breadth groups would be ineffective in destroying local similarity. From Table 4, we conclude that transcriptional breadth profile per se does not importantly underpin local similarity in K_4.

Might the lower K_4 of housekeeping genes likely to be due to transcription-coupled repair reducing the effective mutation rate of germline-expressed genes? In addition to the failure of a breadth-mediated model to explain local similarity, there are at least two other reasons why this conclusion cannot be accepted at face value. First, it has been suggested that this correlation is a secondary effect caused by the K_A-K_4 correlation, as broadly expressed proteins are known to evolve at slower rates (Duret and Mouchiroud 2000; Williams and Hurst 2002). If in some part K_A drives K_4, by whatever mechanism, then one would need to correct for this. Indeed, when we examined residuals of K_4 from a regression on K_A, the breadth-K_4 correlation disappeared (for each data set; data not shown). Correspondingly, the difference between housekeeping (putatively germline-expressed) and tissue-specific genes was much reduced for the residuals (Table 2).

Given that we are unsure whether K_A does drive K_4 (see below), a better test then is to analyze germline-expressed genes alone, and ask if their rate of expression is a good predictor of K_4. To approximate germline transcription rates for each gene, we calculated the median expression rate for all putative housekeeping genes across all (nongermline) tissues with reported expression. This measure does indeed provide a reasonable approximation of the transcription rate in tissues that are not covered by the experiments. To show this, we performed a benchmarking test, by excluding each of the individual tissues in turn, and calculating the median transcription rate only from the other tissues. This measure is highly correlated with the observed transcription rate for the excluded tissue: For each expression measure, Pearson's r is above 0.5 for the vast majority of all tissues (Supplemental Table S1). However, when correlating this housekeeping gene expression rate with K_4, the results are ambiguous. Although we do find a negative correlation for all measures and data sets (except for human SAGE data), this correlation is nonsignificant in most cases (Table 5; to escape a massive reduction in sample size, results are given for human and mouse expression data separately). The correlations become stronger when restricting the analysis to low-GC genes (Table 5); for the same genes, we also find a stronger breadth-K_4 correlation (Table 3). This is con-

Table 3. Correlation Between Synonymous Substitution Rate (K_4) and Expression Breadth

	Human–mouse			Mouse–rat		
Breadth measure[a]	r[b]	P	N	r[b]	P	N
All genes						
EST	−0.203	<0.00001	1247	−0.063	0.00088	2759
SAGE	−0.125	0.00007	1057	−0.088	<0.00001	2579
Microarray	−0.187	0.00003	596	−0.057	0.0237	1592
GC≤0.5						
EST	−0.341	<0.00001	322	−0.165	0.00011	570
SAGE	−0.242	0.00011	260	−0.202	<0.00001	565
Microarray	−0.339	0.00002	148	−0.090	0.098	340

[a]Breadth of expression was averaged over experiments in human and mouse for the human–mouse comparison, and was obtained from mouse only in the mouse–rat comparison.
[b]Pearson's correlation coefficient between expression breadth and K_4. Only nondisparate gene pairs were included.

Table 4. Effect of Expression Breadth on Significance of Local Similarity (ρ) in K_4

Breadth measure[a]	Human–mouse ρ	P_{all}[b]	P_{group}[c]	N	Mouse–rat ρ	P_{all}[b]	P_{group}[c]	N
EST	0.290	<0.0001	<0.0001	933	0.188	<0.0001	<0.0001	2544
SAGE	0.323	<0.0001	<0.0001	732	0.177	<0.0001	<0.0001	2351
Microarray	0.506	<0.0001	<0.0001	320	0.173	<0.0001	<0.0001	1317

[a]Breadth of expression was averaged over experiments in human and mouse for the human–mouse comparison, and was obtained from mouse only in the mouse–rat comparison. Only nondisparate gene pairs were included.
[b]P_{all} is the fraction of equal or greater ρ in datasets obtained by randomly permuting all genes.
[c]P_{group} is the fraction of equal or greater ρ in datasets obtained by randomly permuting genes within classes of similar K_4.

sistent with recent reports of a fixation bias in synonymous substitutions in GC-rich genes (Eyre-Walker 1999; Smith and Eyre-Walker 2001; Duret et al. 2002; Lercher and Hurst 2002a; Lercher et al. 2002a); accordingly, GC-depleted genes may be expected to most accurately reflect mutational biases. Conversely, this result suggests that for the most gene-dense regions, characterized by $GC_4 > 0.5$, systematically varying fixation biases may dominate the transcription-coupled mutational biases examined here.

In sum, we have suggestive evidence that synonymous substitution rates of housekeeping genes may be directly affected by rates of expression, consistent with the action of transcription-coupled repair processes in the mammalian germline (Svejstrup 2002) or with stronger purifying selection acting on the most abundantly expressed genes (Urrutia and Hurst 2003). However, the effect on K_4 is at most weak, and is unlikely to contribute much to the observed patterns of local similarity in K_4.

Alternative Explanations for Local Similarity in K_4

As the transcription-mediated model failed to account for local similarity in K_4, we need to examine alternative hypotheses. The rate of synonymous evolution covaries with regional GC content (Smith and Hurst 1999; Bielawski et al. 2000; Hurst and Williams 2000; Castresana 2002b; Smith et al. 2002; Hardison et al. 2003), although the exact form and strength of this relationship is still a matter of debate, and seems to depend on methodology (Hurst and Williams 2000; Bierne and Eyre-Walker 2003). There is also evidence of a positive correlation between K_4 and meiotic recombination rate (Perry and Ashworth 1999; Lercher and Hurst 2002b; Filatov 2003; Filatov and Gerrard 2003; Hellmann et al. 2003), possibly indicating a mutagenic effect of recombination. GC and recombination rates are known to covary (Eyre-Walker 1993; Fullerton et al. 2001), and it has recently been argued that substitutional GC biases are directly associated with recombination events (Meunier and Duret 2004). Both GC and recombination rate are known to fluctuate systematically over megabase-sized regions, suggesting that regionality in K_4 may be a secondary effect of these variations.

We performed a multiple linear regression of K_4 on recombination rate and on three different GC measures for the human–mouse alignments: GC_4 (GC at fourfold degenerate sites, averaged over the aligned sequences); GC_i (intron GC, averaged over the aligned sequences); and $|\Delta GC_i|$ (absolute difference in intron GC between the aligned sequences). We found that $r^2 = 0.104$ of the variation in K_4 can be explained by these four variables, with all variables contributing significantly (*t*-test, $P < 0.001$ for each variable). When restricting this analysis to nondisparate genes, only GC_4 and recombination rate contribute significantly ($r^2 = 0.077$). We can ask to what extent this covariation contributes to local similarity in K_4, by examining the residuals of the multiple regression. We found that ρ for K_4 is reduced from 0.34 to 0.28 when analyzing the residuals (nondisparate genes: 0.30 to 0.23). Thus, part of the regionality in K_4 can be attributed to effects of GC and recombination rate variation in the human–mouse comparison, even after exclusion of genes exhibiting heterogeneous substitution patterns.

As we were unable to obtain fine-scale recombination rate estimates for rodents, we only tested the influence of the different GC measures on the variation of K_4 in the mouse–rat comparison. In contrast to the above result, the fraction of K_4 variation explained by GC is very low ($r^2 = 0.002$); only GC_4 contributes significantly ($P = 0.016$). This is practically unchanged when restricting the analysis to nondisparate genes ($r^2 = 0.003$). Corre-

Table 5. Correlation Between Expression Rate and K_4 for Putative Housekeeping Genes

Rate measure[a]	Human–mouse r	P	N	Mouse–rat r	P	N
All genes						
SAGE (human)	0.026	0.761	141	—	—	—
Microarray (human)	−0.134	0.079	172	—	—	—
SAGE (mouse)	−0.068	0.388	166	−0.003	0.96	305
Microarray (mouse)	−0.029	0.684	196	−0.050	0.33	391
$GC_4 \leq 0.5$						
SAGE (human)	−0.446	0.006	37	—	—	—
Microarray (human)	−0.038	0.838	32	—	—	—
SAGE (mouse)	−0.401	0.005	47	−0.089	0.43	80
Microarray (mouse)	−0.137	0.319	55	−0.006	0.96	93

[a]Rate measures were not averaged over human and mouse in this table to retain acceptable sample sizes. Only nondisparate gene pairs were included.

spondingly, local similarity was unchanged when analyzing the multiple regression residuals.

One possible reason for a relationship between K_4 and GC content is the hypermutability of CpG dinucleotides. Indeed, it has been suggested that mammalian isochores (regions of varying GC content) are the historical consequence of varying substitution rates at CpG sites (Arndt et al. 2003). To exclude the influence of these sites, we also calculated K_4^{excCpG} after removing all sites contributing to a CpG dinucleotide in any of the aligned sequences. Local similarity is largely unaffected by this (Table 1), suggesting that a substantial fraction of regional variation in the mutation rate is caused by processes at nonCpG sites. Consistent with this supposition, all results reported in Tables 2–7 are essentially unchanged when replacing K_4 with K_4^{excCpG} (Supplemental Tables S2–S6).

It is interesting to note that we found local similarity in K_4 to be strongest for the subset of genes with high GC (human–mouse, $GC_4 > 0.7$: $\rho = 0.473$, $P < 0.0001$, n = 234, nondisparate only). Regions ('isochores') of high GC are also known to exhibit much more small-scale variation in GC than regions of lower GC content (Nekrutenko and Li 2000). It has been suggested that this variation is caused by biased gene conversion at local hotspots of recombination (Meunier and Duret 2004). Thus, it is conceivable that local similarity in K_4 is indeed caused by recombination-induced effects. Our inability to confirm such a relationship may then simply reflect the fact that recombination rate estimates are regional averages (Kong et al. 2002) and are only a poor predictor of the ancestral recombination events that shaped substitution patterns. This view is consistent with a recent study that found a very strong correlation between human recombination rates and the GC bias of recent nucleotide substitutions (Meunier and Duret 2004).

Breadth of Expression Has, at Most, a Weak Effect on Local K_A Similarity

Before we can assess the contribution of expression-mediated effects on K_A, we must first exclude duplicated genes from the analysis (Williams and Hurst 2000; Lercher et al. 2001). As duplicated genes may often be subject to similar selective constraints, their rates of amino acid substitution will be correlated. Duplicated genes often reside close to each other and may thus contribute to a signal of local similarity. We compared the K_A between the two copies of 472 human–mouse duplicate gene pairs, identified by significant sequence similarity (pairwise blast expectation value $E \leq 0.001$), with at most 1 Mb distance between them on a human autosome. As expected, the K_A values of duplicated genes are strongly correlated ($r = 0.54$). Surprisingly, the same protocol showed a much weaker similarity of duplicate genes in the mouse–rat comparison ($r = 0.26$). It is interesting to note that K_A values are still correlated between very distantly related sequences: When sorting neighboring gene pairs (those within 1 Mb of each other) into subsets of 100 pairs according to pairwise BLAST score, we found that the correlation between K_A values was significantly increased for all sets with expectation values $E < 0.01$ (Lercher et al. 2002b). To test the effect of these correlations on the local K_A similarity, we recalculated ρ, this time excluding all putative duplicates (Table 6). ρ was reduced by ~50% in both the human–mouse and mouse–rat comparisons. Local similarity was still significant in the human–mouse comparison. However, in the mouse–rat comparison, local similarity in K_A became nonsignificant after strict removal of duplicate genes. In the mouse–rat comparison, local K_A similarity (before duplicate exclusion) was enhanced when restricting the analysis to low-GC genes, and was strengthened by further excluding genes with disparate substitution patterns ($\rho = 0.21$). For this subclass of genes, local similarity remained significant even after exclusion of duplicates ($\rho = 0.17$, $P = 0.0023$).

Table 6. Local Similarity in Nonsynonymous Substitution Rate K_A, Including All Genes or Excluding Duplicate Genes

	Human–mouse			Mouse–rat		
	ρ	*P*	*N*	ρ	*P*	*N*
All genes	0.126	<0.0001	4596	0.055	0.0018	4116
Nonduplicates[a]	0.065	0.0002	4515	0.022	0.087	3995

[a]Genes were excluded from the focal average if they exhibited significant sequence similarity to the focal gene (BLAST expectation value $E < 0.02$; Lercher et al. 2001).

We then proceeded to analyze the influence of gene expression on K_A. In agreement with previous studies (Duret and Mouchiroud 2000; Williams and Hurst 2002), our analysis confirmed a significant negative correlation between different measures of expression breadth and the rate of protein evolution, K_A (Table 8).Depending on the data sets used, between 2% and 20% of the variation in K_A can be predicted by a gene's expression breadth. Although these numbers appear relatively low compared to Figure 1 in Duret and Mouchiroud (2000), it must be kept in mind that the latter study grouped genes according to similar expression breadth, and thereby filtered out a large proportion of additional variation. When comparing different expression assays, we found that the strongest correlations are consistently seen with EST data; this may be a simple consequence of the large number of tissues for which EST data are available. Further, it is striking that the K_A-breadth correlation is enhanced when restricting the analysis to low-GC genes (Table 8). This is reminiscent of a similar effect for K_4 (Table 3).

Next, we analyzed local similarity (ρ, excluding duplicate genes) for all genes with valid measures of expression and K_A, assessing statistical significance by comparison to 10,000 randomized data sets. To test whether regional variation in expression breadth is responsible for part of the local similarity in K_A, we repeated the randomization procedure, this time permuting gene positions only between genes with similar breadth of expression (Table 9). The local similarity measure ρ in Table 9 is generally low, probably due to the sample-size reductions associated with the expression data sets. Controlling for expression breadth has practically no influence on the estimated statistical significance of ρ ($P_{all} \approx P_{group}$). We conclude that, at most, only a small part of local similarity in K_A can be explained by the covariation with expression breadth, and alternative explanations must be sought.

Alternative Explanations for Local Similarity in K_A

As commonly described (Li 1997; Smith and Hurst 1999), a sizeable part of the variation in the protein sequence rate of evolution (K_A) is predicted by K_4, the substitution rate at fourfold degenerate sites (human–mouse: $r = 0.39$, $P < 0.00001$, n = 4726; mouse–rat: $r = 0.23$, $P < 0.00001$, n = 4092; but see also Bielawski et al. 2000). This covariation has been attributed, in part, to correlated (but not necessarily simultaneous) substitutions between neighboring synonymous and nonsynonymous sites (tandem substitutions; Smith and Hurst 1999; Duret and Mouchiroud 2000). When, for example, mouse–rat genes with no tandem substitutions are analyzed, there is no K_A-K_4 correlation (Smith and Hurst 1999). When following the method of Duret and Mouchiroud (2000), by excluding all fourfold degenerate changes neighboring a substitution at the first site of the next codon, we find

Table 7. Effect of K_4 on Significance of Local Similarity in K_A (Duplicate Genes Excluded)

	Human–mouse				Mouse–rat			
	ρ	P_{all}[a]	P_{group}[b]	*N*	ρ	P_{all}[a]	P_{group}[b]	*N*
All genes	0.053	0.0008	0.046	4191	0.026	0.063	0.093	3787
Nondisparate	0.059	0.025	0.21	1312	0.039	0.018	0.029	3264

[a] P_{all} is the number of equal or greater ρ in datasets obtained by randomly permuting all genes.
[b] P_{group} is the number of equal or greater ρ in datasets obtained by randomly permuting genes within classes of similar K_4.

that the correlation is substantially reduced (human–mouse: $r = 0.13$, $P < 0.00001$, n = 4749; mouse–rat: $r = 0.13$, $P < 0.00001$, n = 4085). However, this could be an overestimate of the impact of tandem substitutions: even if neighboring substitutions occur independent of each other, fast-evolving genes will have far more tandem substitutions than slowly evolving genes ($\sim K_A \times K_4$). Thus, we will take out disproportionately more synonymous substitutions for fast-evolving genes than for slowly evolving genes, which of course reduces the correlation between K_A and K_4.

To get an unbiased estimate of the underlying mutation rate excluding any potential tandem substitution effects, we must adjust not just the number of substitutions, but also the number of fourfold degenerate sites that are used to calculate substitutions per site. Thus, we recalculated K_4, this time excluding all fourfold degenerate sites (with or without substitutions) that were followed by a codon with a substitution at its first site. This new K_4' was still correlated with K_A (human–mouse: $r = 0.28$, $P < 0.00001$, n = 4725; mouse–rat: $r = 0.18$, $P < 0.00001$, n = 4064).

In sum, although tandem substitution biases appear to exist [K_4' is on average 3% (human–mouse) or 1.6% (mouse–rat) lower than K_4], these biases cannot fully account for the K_A-K_4 correlation. This finding is consistent with a number of recent analyses, which, in contrast to early reports (Averof et al. 2000), find evidence for only a very low rate of doublet mutations (Silva and Kondrashov 2002; Kondrashov 2003; Smith et al. 2003). In further support of our conclusion that the K_A-K_4 correlation is not due to mechanistic coupling of substitutions of neighboring sites, we also found a significant positive correlation between K_A and K_i (the substitution rate within introns of the same gene, $r_{Spearman} = 0.248$, $P = 0.005$, n = 127).

The strong K_A-K_4 correlation demonstrated above suggests that much of the local similarity in K_A may be explained by the local similarity in K_4. To test this hypothesis, we calculated ρ values for K_A across all genes with valid estimates for K_A and K_4, excluding duplicate genes from the focal averages. Statistical significance was estimated by comparison of ρ to 10,000 data sets obtained by randomly permuting the positions of all genes (P_{all}, Table 7). We then tested for the effect of K_4, by repeating the randomization procedure, this time permuting only gene positions within classes of similar K_4 (P_{group}, Table 7). For the human–mouse comparison, significance of the ρ value was markedly reduced, and became marginally significant (all genes; nonsignificant when ρ was calculated as Spearman's rank correlation coefficient) or nonsignificant (nondisparate genes or GC ≤ 0.5). Thus, after the exclusion of duplicate genes, the majority of local K_A similarity in the human–mouse comparison can be attributed to the genes' synonymous substitution rates.

Consistent with an earlier analysis (Williams and Hurst 2000), we obtained a markedly different result in the mouse–rat comparison (Table 7). Significance of the ρ value hardly depends on randomization protocol ($P_{all} \approx P_{group}$), suggesting that the underlying mutation rate contributes very little to the local similarity among rodents. In fact, local similarity in K_A was nonsignificant for the mouse–rat comparison after removal of duplicate genes (Table 6). What causes this difference between the two species comparisons? If we accept the notion that a coupling of K_A and K_4 is responsible for the local similarity observed in the human–mouse comparison, this suggests that the K_A-K_4 coupling is reduced in rodents. One possible reason may be that the effective population sizes of rodents are larger, and thus fewer amino acid substitutions are effectively neutral; however, further analyses are necessary to resolve this issue.

A Model for the Strength of Local Similarity in K_A

In apparent contrast to our finding that transcription does not have a significant role in local K_A similarity, a previous analysis of mouse–rat orthologs reported that local similarity in K_A is most pronounced in the vicinity of narrowly expressed (tissue-specific) genes (Williams and Hurst 2002). This we have confirmed for our human–mouse data set. When analyzing separately each of five breadth classes, we found significant local similarity only for the most narrowly expressed genes (excluding duplicate genes: ρ = 0.084, $P = 0.010$ including Bonferroni correction for multiple tests, n = 1241). The local similarity estimated for this subset of

Table 8. Correlation Between Expression Breadth and Nonsynonymous Substitution Rate, K_A

	Human–mouse			Mouse–rat		
Breadth measure	*r*	*P*	*N*	*r*	*P*	*N*
All genes						
EST	−0.286	<0.00001	3451	−0.204	<0.00001	3305
SAGE	−0.197	<0.00001	2925	−0.158	<0.00001	3075
Microarray	−0.242	<0.00001	1624	−0.133	<0.00001	1909
Nondisparate & GC≤0.5						
EST	−0.463	<0.00001	322	−0.341	<0.00001	570
SAGE	−0.276	<0.00001	260	−0.200	0.00001	565
Microarray	−0.363	<0.00001	148	−0.225	0.00003	340

Table 9. Effect of Expression Breadth on Significance of Local Similarity in K_A (Duplicate Genes Excluded)

	Human–mouse				Mouse–rat			
Breadth measure[a]	ρ	P_{all}[b]	P_{group}[c]	*N*	ρ	P_{all}[b]	P_{group}[c]	*N*
EST	0.047	0.0072	0.016	3115	0.023	0.10	0.069	3029
SAGE	0.041	0.032	0.044	2572	0.041	0.024	0.020	2740
Microarray	−0.037	0.90	0.92	1209	0.038	0.071	0.049	1559

[a]Breadth of expression was averaged over experiments in human and mouse for the human–mouse comparison, and was obtained from mouse only in the mouse–rat comparison.
[b]P_{all} is the number of equal or greater ρ in datasets obtained by randomly permuting all genes.
[c]P_{group} is the number of equal or greater ρ in datasets obtained by randomly permuting genes within classes of similar expression breadth.

putative tissue-specific genes is actually higher than that estimated for the total data set (ρ = 0.065). Does this imply a direct coupling of gene expression with regionality in K_A? We wish to suggest that this might not necessarily be so.

Consider a simple model supposing that the local similarity in K_A (excluding duplicate gene effects) is driven by the local similarity in the mutation rate (Malcom et al. 2003). Let us further assume that K_4 is somehow coupled to the mutation rate, which as noted above need not be true. If one subsamples any given group of genes, the extent to which one will detect local similarity will then depend on the extent to which, within the subsample, K_A is coupled to K_4. Consider now two extremes: (1) a set of proteins that evolve neutrally ($K_A = K_4$), and (2) another set under extreme purifying selection ($K_A = 0$ for all). In the former, K_A and K_4 are perfectly coupled; in the latter there is no coupling. More generally, when we select subsamples under different degrees of purifying selection, then lower selection pressures will correspond to a higher proportion of effectively neutral amino acid substitutions, and thus to stronger K_A-K_4 coupling. If narrowly expressed genes are under weaker purifying selection (as appears to be the case, see above), then we expect a tighter coupling of K_A and K_4, and consequently a stronger signal of local similarity when only these genes are analyzed. Our finding that the K_A-K_4 correlation cannot be accounted for by tandem mutations is of importance for this interpretation, as otherwise one might suppose that K_A drives K_4, not the other way around.

As expected from our model, the K_A-K_4 coupling is strongest in tissue-specific genes: $r = 0.42$, compared to $r = 0.39$ when including all genes (human SAGE data). This observation also strengthens the notion that the coupling is caused by a fraction of effectively neutrally evolving amino acid positions. If alternatively the coupling was due to similar selection pressures on both nonsynonymous and synonymous sites, then the coupling should be stronger when considering the full range of expression profiles.

We applied two additional tests for our model. We first asked whether the observed strength of the K_A-K_4 coupling in randomly drawn subsets of our data predicts the strength of local similarity found within the subset: this is indeed the case ($r_{Spearman} = 0.083$, $P = 0.004$, from examining the dependence of ρ on the correlation coefficient r between K_A and K_4, for 1000 randomly drawn subsets of 1000 genes). Secondly, we compared the quarter of our data set with the highest K_A/K_4 to the quarter exhibiting the lowest K_A/K_4; these groups putatively correspond to genes under low and high selective pressures, respectively. As expected, we found stronger similarity in the genes with higher K_A/K_4 (excluding duplicate genes: ρ = 0.132 vs. 0.084); this group also exhibits stronger K_A-K_4 coupling ($r = 0.62$ vs. 0.51).

Thus, the stronger local similarity for narrowly expressed genes might be explained as a consequence of the stronger K_A-K_4 coupling, which in turn is due to a fraction of sites that evolve effectively neutrally. There is, however, at least one problem with the null model, this being that the local similarity in K_4 extends over many megabases (Smith and Lercher 2002), compared to less than 3 Mb for local similarity in K_A (Lercher et al. 2001). The local similarity in K_4 decreases, however, with increasing distance. It is thus possible that at larger distances secondary effects on K_A are present, but are too small to be detected.

In sum, we have demonstrated that transcription has little to do with establishing local similarity in rates of evolution. Local similarity in K_A appears to be largely due to tandemly duplicated genes, and to the coupling of K_A to the mutation rate for sites encoding amino acids that evolve neutrally. In turn, the local similarity in K_4 may be largely due to recombination-associated effects.

METHODS

Accession numbers of orthologs and associated data used for the analyses are available as online Supplemental data.

Orthologous Coding Sequence Identification

We obtained lists of putative human–mouse and mouse–rat orthologous genes, identified through reciprocal best BLAST hits, from Ensembl (http://www.ensembl.org; Hubbard et al. 2002). If a gene in one species matched more than one gene in the other species, indicating a lineage-specific gene duplication, it was excluded from further analysis. This resulted in primary data sets of 13,015 (human–mouse) and 12,637 (mouse–rat) orthologous gene pairs. We excluded human transcripts without known position on the UCSC November 2002 genome assembly (http://genome.ucsc.edu), and mouse genes without known position on the Ensembl map. As evolutionary forces affecting genes located on the sex chromosomes differ from those affecting autosomal genes, we restrict our analyses to genes located on human or mouse autosomes.

For each gene, we then downloaded all transcripts from Ensembl. We excluded those transcripts where the coding sequence lacked a valid start or stop codon. For each orthologous gene pair, we selected matching transcripts. This was done under the assumption that a large proportion of genes is alternatively spliced, and that two transcripts corresponding to analogous splice forms should have similar lengths. We first searched for the longest pair of transcripts with a length difference of at most 1%; if no transcript pair fulfilled this criterion, we selected the transcript pair most similar in length. If all transcript pairs differed by more than 5% in their length, we discarded the gene. This procedure resulted in sets of 5212 (human–mouse) and 4442 (mouse–rat) transcript pairs, where transcript pairs will generally correspond to analogous splice forms. For the human–mouse comparison, distances were measured between transcription midpoints of genes on the human UCSC November 2002 assem-

bly; for the mouse–rat comparison, we used the Ensembl mouse map (build 30).

Nucleotide Alignments and Evolutionary Distances

Transcript coding sequences were first translated to amino acid sequences. These were aligned using Clustalw (Thompson et al. 1994) with default settings. The amino acid alignments were then used as templates to align the nucleotides. We calculated nonsynonymous distances (K_A) using the method of Li (1993) and the Kimura two-parameter model. The neutral substitution rate was estimated from the distance at fourfold degenerate sites (K_4), using only codons with no changes at other sites. These rates were corrected for multiple hits with a model that accounts for compositional biases and for substitution pattern differences (disparity) between the two sequences (Tamura and Kumar 2002). We defined gene classes of similar K_4 by dividing the ranked data set into 10 equally sized groups.

For intronic substitution rates, only introns located between coding exons were analyzed. Two methods were used for aligning introns: manually (by-eye) and MCALIGN, a stochastic maximum likelihood-based (ML) program that incorporates a Monte Carlo algorithm (http://homepages.ed.ac.uk/eang33/mcinstructions.html). We executed the program using the rodent intron parameters provided. Seven large introns (>7 kb) proved too difficult to align. Our final intron data set consisted of 136 orthologous genes possessing 560 introns. For further details see Chamary and Hurst (2004).

Within introns we excluded the first and last 20 base pairs, as these appear to be subject to purifying selection (Majewski and Ott 2002; Chamary and Hurst 2004). To obtain genic K_i values, intronic substitution rates were weighted according to the number of bases compared per individual intron alignment (Smith and Hurst 1998). The indel rate, K_{indel}, was calculated as the total number of indels per base pair of the alignment. After estimating the K_i per intron by the two alignment methods (manual and from the ML protocol), we defined a conservative set and a liberal set. For any given intron, these two sets contain the alignment that yields the lower or higher K_i respectively, providing two estimates for the rate of evolution of each intron. We here report results only for the conservative set; very similar results were obtained for the liberal set, as well as for an additional set that consists of only slow-evolving regions (data not shown; Castresana 2000; http://www1.imim.es/~castresa/Gblocks/Gblocks.html).

GC Content and Recombination Rates

For each gene, the guanine + cytosine content at fourfold degenerate sites, GC_4, was averaged over the two aligned coding sequences. We further calculated intron GC for each transcript from the compositional difference between mRNA and exon sequences (downloaded from Ensembl), as $GC_{intron} = (GC_{mRNA} \times length_{mRNA} - GC_{exons} \times length_{exons}) / (length_{mRNA} - length_{exons})$. From this, we calculated GC_i^{avg}, the average over the two aligned transcripts, and GC_i^{diff}, the absolute difference between the two transcripts. Recombination rate estimates for humans were obtained from the UCSC genome browser (http://genome.ucsc.edu), and are based on the deCODE data (Kong et al. 2002).

Expression Breadth and Rate

EST Data

Human and mouse Ensembl genes were mapped to NCBI UniGene clusters (Schuler et al. 1996; UniGene build 161, obtained from NCBI at ftp://ncbi.nlm.nih.gov/repository/UniGene) via RefSeq sequence IDs. Only unambiguous pairings were retained. dbEST library accessions for all ESTs mapping to these clusters were extracted from UniGene. For each library mapping to at least 50 UniGene clusters, the associated tissue type was obtained from dbEST annotation (http://ncbi.nlm.nih.gov/dbEST). We kept only libraries based on well defined, nondisease tissue types. Libraries representing the same tissue type were joined, and tissues matching <500 Ensembl genes were excluded. This resulted in a data set containing 14,559 genes expressed in at least one out of 55 tissue types for humans, and in a second set containing 11,418 genes expressed in at least one out of 49 tissue types for mouse. For each gene, breadth of expression was estimated as the fraction of tissues with an observed EST.

SAGE Data

Serial Analysis of Gene Expression (SAGE) data (Velculescu et al. 1995) was obtained from SAGEmap (Lash et al. 2000) at NCBI (ftp://ncbi.nlm.nih.gov/pub/sage). The data sets were curated to avoid possible GC biases in SAGE libraries, following the approach of Margulies et al. (2001), by removing libraries with mean tag GC > 0.5. The resulting SAGE tag/tissue data sets were based on 40 libraries representing 19 nondisease tissues (human), and on 23 libraries representing nine nondisease tissues (mouse). Tag counts for each data set were converted to relative values (cpm, counts per million) after joining all libraries representing the same tissue type. If tags were found only once in one tissue type, we discarded the observation as a likely sequencing error. These data sets were cross-linked to the mRNA sequences in RefSeq (ftp://ncbi.nlm.nih.gov/refseq), by extracting the 3′-most NlaIII and Sau3A SAGE tags for each human and mouse mRNA. These were then cross-linked to Ensembl genes. We disregarded all tags mapping to more than one Ensembl gene, and excluded the associated genes from further analysis. If several tags mapped to the same gene (representing alternative splice forms), we used maximum cpm in each tissue. In human, we obtained reliable SAGE tags for 11,507 genes, with 7285 expressed in at least one tissue. In mouse, we collected expression data for 10,480 genes, of which 5016 were expressed in at least one tissue. For each gene, we calculated breadth of expression as the fraction of tissues with cpm > 0. Rate of expression was defined as cpm in each tissue.

Microarray Data

Normalized microarray expression data based on Affymetrix chips for 7315 human and 5971 mouse genes were obtained from Su et al. (2002). Human data were sorted into 28 nonredundant tissue types, encompassing 63 replicate hybridizations. Mouse data for 45 tissue types were based on 98 replicate hybridizations. For each tissue, the mRNA expression level (termed 'expression rate' to be consistent with SAGE terminology) was estimated as the mean across replicates. Because there is no unambiguous way to distinguish expressed from nonexpressed data in this type of experiment, we based our breadth measure on observed expression rates, as breadth = (average mRNA expression level across tissues) / (maximum mRNA expression level across tissues). Breadth was set to 0 if the mRNA expression level was <50 in all tissues; this low level could be chosen because our method effectively smoothes out experimental error by joining information across tissues. We also tried other breadth measures (such as defining genes with level < 100 as nonexpressed, and genes with level > 200 as expressed; Su et al. 2002), with very similar results (data not shown).

Definition of Breadth Classes

For the analysis in similar breadth classes, we subdivided the data set into classes of width 0.05. As the highest breadth classes contain the lowest numbers of genes, we further joined these until the highest class contained at least 20 genes. For further analysis of subgroups of genes (nondisparate, low-GC), we joined neighboring classes until each class contained at least 10 genes.

All expression assays are biased against genes expressed at low levels, which may not be detected unless very high numbers of ESTs, SAGE tags, or replicate hybridizations are analyzed. For this reason, we expect many cases where absence of gene expression is wrongly inferred from the data. Accordingly, we must allow for false negatives in some tissues when selecting putative housekeeping genes. For microarray data, we treated all mRNA expression levels <100 as nonexpressed, all >200 as expressed, and all other as unknown (Su et al. 2002). We conservatively labeled those genes without nonexpressed tissues as putative

housekeeping genes (human: 8.5% of genes; mouse: 11.1%). Based on this analysis, we estimated that at least 10% of all genes should perform housekeeping functions. For EST and SAGE data, we selected corresponding thresholds of observed expression breadth for putative housekeeping genes: human SAGE, 0.7 (7.6% of genes); mouse SAGE, 0.5 (9.0%); human EST, 0.52 (9.7%); mouse EST, 0.6 (7.9%). In all assays, putative tissue-specific genes were those with reported expression in 0 or 1 of our tissues.

For the analysis of local similarity within different breadth classes, we classified genes according to the number of tissues with expression reported in human SAGE experiments: 0–1 (tissue-specific), 2–4, 5–8, 9–13, and 14–19 (broadly expressed).

Focal Average Correlation (ρ) and Statistics

In a modification of the method of Lercher et al. (2001), we first calculated focal averages of substitution rates. For each gene, we identified all other genes within 1 Mb along the chromosome, and calculated their mean substitution rate. On average, each gene had eight and nine such neighbors in the human–mouse and mouse–rat comparisons, respectively. When including only nondisparately evolving gene pairs, this was reduced to four and eight genes, respectively. When restricting the analysis to a certain class of genes, we included only focal genes and only neighboring genes within the same class. We then calculated Pearson's correlation coefficient across all data pairs consisting of the rate of a gene and the corresponding focal average; we denote this focal average correlation by ρ. A randomization protocol was employed to assess statistical significance. Gene positions were randomly permuted $N_0 = 10{,}000$ times, and ρ_{rand} was calculated for each random data set. n_P was the number of random data sets for which $\rho_{rand} \geq \rho$. From this, we estimated $P = (n_P + 1) / (N_0 + 1)$. This protocol maintains the original data structure; in particular, it leaves the distribution of neighbors unchanged.

To assess the correlation of K_i or K_{indel} values across different introns of the same gene, we similarly chose a focal intron and defined the focal average as the mean over all other introns in that gene. We then calculated Spearman's correlation coefficient (ρ) for data pairs consisting of the focal introns and their focal average. Statistical significance was estimated as above.

Throughout the paper, r denotes Pearson's correlation coefficient. Statistical significance was estimated by randomly permuting the values in one of the two columns. Repeating this $N_0 = 100{,}000$ times, we counted the number of times n_P when $r_{rand}^2 \geq r^2$, where r_{rand} is the correlation coefficient for the randomized data. From this, we estimate a two-sided $P = (n_P + 1) / (N_0 + 1)$. All correlation and focal average analyzes were also performed for Spearman's rank correlation coefficient, with very similar results (data not shown).

Before calculating linear regressions, we log-transformed values for substitution rates and GC measures.

ACKNOWLEDGMENTS

We thank Laurent Duret and Itai Yanai for helpful discussions, and four anonymous reviewers for comments on the manuscript. M.J.L. acknowledges financial support by The Wellcome Trust and the Royal Society. J.V.C. and L.D.H. are funded by the UK Biotechnology and Biological Sciences Research Council.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Aguilera, A. 2002. The connection between transcription and genomic instability. *EMBO J.* **21:** 195–201.

Arndt, P.F., Petrov, D.A., and Hwa, T. 2003. Distinct changes of genomic biases in nucleotide substitution at the time of mammalian radiation. *Mol. Biol. Evol.* **20:** 1887–1896.

Averof, M., Rokas, A., Wolfe, K.H., and Sharp, P.M. 2000. Evidence for a high frequency of simultaneous double-nucleotide substitutions. *Science* **287:** 1283–1286.

Bielawski, J.P., Dunn, K.A., and Yang, Z. 2000. Rates of nucleotide substitution and mammalian nuclear gene evolution: Approximate and maximum-likelihood methods lead to different conclusions. *Genetics* **156:** 1299–1308.

Bierne, N. and Eyre-Walker, A. 2003. The problem of counting sites in the estimation of the synonymous and nonsynonymous substitution rates. Implications for the correlation between the synonymous substitution rate and codon usage bias. *Genetics* **165:** 1587–1597.

Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M.C., van Asperen, R., Boon, K., Voute, P.A., et al. 2001. The human transcriptome map: Clustering of highly expressed genes in chromosomal domains. *Science* **291:** 1289–1292.

Casane, D., Boissinot, S., Chang, B.H.J., Shimmin, L.C., and Li, W.H. 1997. Mutation pattern variation among regions of the primate genome. *J. Mol. Evol.* **45:** 216–226.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17:** 540–552.

———. 2002a. Estimation of genetic distances from human and mouse introns. *Genome Biol.* **3:** RESEARCH0028.

———. 2002b. Genes on human chromosome 19 show extreme divergence from the mouse orthologs and a high GC content. *Nucleic Acids Res.* **30:** 1751–1756.

Chamary, J.V. and Hurst, L.D. 2004. Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: Evidence for selectively driven codon usage. *Mol. Biol. Evol.* (in press).

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

Duret, L. and Mouchiroud, D. 2000. Determinants of substitution rates in mammalian genes: Expression pattern affects selection intensity but not mutation rate. *Mol. Biol. Evol.* **17:** 68–74.

Duret, L., Semon, M., Piganeau, G., Mouchiroud, D., and Galtier, N. 2002. Vanishing GC-rich isochores in mammalian genomes. *Genetics* **162:** 1837–1847.

Ebersberger, I., Metzler, D., Schwarz, C., and Paabo, S. 2002. Genomewide comparison of DNA sequences between humans and chimpanzees. *Am. J. Hum. Genet.* **70:** 1490–1497.

Eyre-Walker, A. 1993. Recombination and mammalian genome evolution. *Proc. R. Soc. London Ser. B* **252:** 237–243.

———. 1999. Evidence of selection on silent site base composition in mammals: Potential implications for the evolution of isochores and junk DNA. *Genetics* **152:** 675–683.

Filatov, D.A. 2003. A gradient of silent substitution rate in the human pseudoautosomal region. *Mol. Biol. Evol.* **21:** 410–417.

Filatov, D.A. and Gerrard, D.T. 2003. High mutation rates in human and ape pseudoautosomal genes. *Gene* **317:** 67–77.

Fullerton, S.M., Bernardo Carvalho, A., and Clark, A.G. 2001. Local rates of recombination are positively correlated with GC content in the human genome. *Mol. Biol. Evol.* **18:** 1139–1142.

Green, P., Ewing, B., Miller, W., Thomas, P.J., Nc, N., and Green, E.D. 2003. Transcription-associated mutational asymmetry in mammalian evolution. *Nat. Genet.* **33:** 514–517.

Hardison, R.C., Roskin, K.M., Yang, S., Diekhans, M., Kent, W.J., Weber, R., Elnitski, L., Li, J., O'Connor, M., Kolbe, D., et al. 2003. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res.* **13:** 13–26.

Hastings, K.E.M. 1996. Strong evolutionary conservation of broadly expressed protein isoforms in the troponin I gene family and other vertebrate gene families. *J. Mol. Evol.* **42:** 631–640.

Hellmann, I., Ebersberger, I., Ptak, S.E., Paabo, S., and Przeworski, M. 2003. A neutral explanation for the correlation of diversity with recombination rates in humans. *Am. J. Hum. Genet.* **72:** 1527–1535.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., et al. 2002. The Ensembl genome database project. *Nucleic Acids Res.* **30:** 38–41.

Hughes, A.L. and Yeager, M. 1997. Comparative evolutionary rates of introns and exons in murine rodents. *J. Mol. Evol.* **45:** 125–130.

Huminiecki, L., Lloyd, A.T., and Wolfe, K.H. 2003. Congruence of tissue expression profiles from Gene Expression Atlas, SAGEmap and TissueInfo databases. *BMC Genomics* **4:** 31.

Hurst, L.D. and Eyre-Walker, A. 2000. Evolutionary genomics: Reading the bands. *Bioessays* **22:** 105–107.

Hurst, L.D. and Williams, E.J.B. 2000. Covariation of GC content and the silent site substitution rate in rodents: Implications for methodology and for the evolution of isochores. *Gene* **261:** 107–114.

Kondrashov, A.S. 2003. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Hum. Mut.* **21:** 12–27.

Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., et al. 2002. A high-resolution recombination map of the human genome. *Nat. Genet.* **31:** 241–247.
Kumar, S. and Gadagkar, S.R. 2001. Disparity index: A simple statistic to measure and test the homogeneity of substitution patterns between molecular sequences. *Genetics* **158:** 1321–1327.
Kumar, S. and Subramanian, S. 2002. Mutation rates in mammalian genomes. *Proc. Natl. Acad. Sci.* **99:** 803–808.
Lash, A.E., Tolstoshev, C.M., Wagner, L., Schuler, G.D., Strausberg, R.L., Riggins, G.J., and Altschul, S.F. 2000. SAGEmap: A public gene expression resource. *Genome Res.* **10:** 1051–1060.
Lercher, M.J. and Hurst, L.D. 2002a. Can mutation or fixation biases explain the allele frequency distribution of human single nucleotide polymorphisms (SNPs)? *Gene* **300:** 53–58.
———. 2002b. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet.* **18:** 337–340.
Lercher, M.J., Williams, E.J.B., and Hurst, L.D. 2001. Local similarity in evolutionary rates extends over whole chromosomes in human–rodent and mouse–rat comparisons: Implications for understanding the mechanistic basis of the male mutation bias. *Mol. Biol. Evol.* **18:** 2032–2039.
Lercher, M.J., Smith, N.G., Eyre-Walker, A., and Hurst, L.D. 2002a. The evolution of isochores: Evidence from SNP frequency distributions. *Genetics* **162:** 1805–1810.
Lercher, M.J., Urrutia, A.O., and Hurst, L.D. 2002b. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat. Genet.* **31:** 180–183.
Lercher, M.J., Urrutia, A.O., Pavlicek, A., and Hurst, L.D. 2003. A unification of mosaic structures in the human genome. *Hum. Mol. Genet* **12:** 2411–2415.
Li, W.-H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.* **36:** 96–99.
———. 1997. *Molecular evolution.* Sinauer, Sunderland, Mass.
Majewski, J. and Ott, J. 2002. Distribution and characterization of regulatory elements in the human genome. *Genome Res.* **12:** 1827–1836.
Malcom, C.M., Wyckoff, G.J., and Lahn, B.T. 2003. Genic mutation rates in mammals: Local similarity, chromosomal heterogeneity, and X-versus-autosome disparity. *Mol. Biol. Evol.* **20:** 1633–1641.
Margulies, E.H., Kardia, S.L., and Innis, J.W. 2001. Identification and prevention of a GC content bias in SAGE libraries. *Nucleic Acids Res.* **29:** e60.
Matassi, G., Sharp, P.M., and Gautier, C. 1999. Chromosomal location effects on gene sequence evolution in mammals. *Curr. Biol.* **9:** 786–791.
Mellon, I., Spivak, G., and Hanawalt, P.C. 1987. Selective removal of transcription-blocking DNA damage from the transcribed strand of the mammalian Dhfr gene. *Cell* **51:** 241–249.
Meunier, J. and Duret, L. 2004. Recombination drives the evolution of GC-content in the human genome. *Mol. Biol. Evol.* (in press).
Nachman, M.W. and Crowell, S.L. 2000. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156:** 297–304.
Nekrutenko, A. and Li, W.H. 2000. Assessment of compositional heterogeneity within and between eukaryotic genomes. *Genome Res.* **10:** 1986–1995.
Ogata, H., Fujibuchi, W., and Kanehisa, M. 1996. The size differences among mammalian introns are due to the accumulation of small deletions. *FEBS Lett.* **390:** 99–103.
Perry, J. and Ashworth, A. 1999. Evolutionary rate of a gene affected by chromosomal position. *Curr. Biol.* **9:** 987–989.
Schuler, G.D., Boguski, M.S., Stewart, E.A., Stein, L.D., Gyapay, G., Rice, K., White, R.E., Rodriguez-Tome, P., Aggarwal, A., Bajorek, E., et al. 1996. A gene map of the human genome. *Science* **274:** 540–546.
Selby, C.P. and Sancar, A. 1993. Transcription-repair coupling and mutation frequency decline. *J. Bacteriol.* **175:** 7509–7514.
Silva, J.C. and Kondrashov, A.S. 2002. Patterns in spontaneous mutation revealed by human–baboon sequence comparison. *Trends Genet.* **18:** 544–547.
Smith, N.G.C. and Eyre-Walker, A. 2001. Synonymous codon bias is not caused by mutation bias in G+C- rich genes in humans. *Mol. Biol. Evol.* **18:** 982–986.
Smith, N.G.C. and Hurst, L.D. 1998. Sensitivity of patterns of molecular evolution to alterations in methodology: A critique of Hughes and Yeager. *J. Mol. Evol.* **47:** 493–500.
———. 1999. The effect of tandem substitutions on the correlation between synonymous and nonsynonymous rates in rodents. *Genetics* **153:** 1395–1402.
Smith, N.G. and Lercher, M.J. 2002. Regional similarities in polymorphism in the human genome extend over many megabases. *Trends Genet.* **18:** 281–283.
Smith, N.G.C., Webster, M.T., and Ellegren, H. 2002. Deterministic mutation rate variation in the human genome. *Genome Res.* **12:** 1350–1356.
———. 2003. A low rate of simultaneous double-nucleotide mutations in primates. *Mol. Biol. Evol.* **20:** 47–53.
Su, A.I., Cooke, M.P., Ching, K.A., Hakak, Y., Walker, J.R., Wiltshire, T., Orth, A.P., Vega, R.G., Sapinoso, L.M., Moqrich, A., et al. 2002. Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl. Acad. Sci.* **99:** 4465–4470.
Svejstrup, J.Q. 2002. Mechanisms of transcription-coupled DNA repair. *Nat. Rev. Mol. Cell Biol.* **3:** 21–29.
Tamura, K. and Kumar, S. 2002. Evolutionary distance estimation under heterogeneous substitution pattern among lineages. *Mol. Biol. Evol.* **19:** 1727–1736.
Tamura, K. and Nei, M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial-DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10:** 512–526.
Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. Clustalw—Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.
Urrutia, A.O. and Hurst, L.D. 2003. The signature of selection mediated by expression on human genes. *Genome Res.* **13:** 2260–2264.
van Gool, A.J., van der Horst, G., Citterio, E., and Hoeijmakers, J.H.J. 1997. Cockayne syndrome: Defective repair of transcription? *EMBO J.* **16:** 4155–4162.
Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W. 1995. Serial analysis of gene expression. *Science* **270:** 484–487.
Versteeg, R., van Schaik, B.D.C., van Batenburg, M.F., Roos, M., Monajemi, R., Caron, H., Bussemaker, H.J., and van Kampen, A.H.C. 2003. The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res.* **13:** 1998–2004.
Williams, E.J.B. and Hurst, L.D. 2000. The proteins of linked genes evolve at similar rates. *Nature* **407:** 900–903.
———. 2002. Clustering of tissue-specific genes underlies much of the similarity in rates of protein evolution of linked genes. *J. Mol. Evol.* **54:** 511–518.
Yi, S.J., Ellsworth, D.L., and Li, W.H. 2002. Slow molecular clocks in Old World monkeys, apes, and humans. *Mol. Biol. Evol.* **19:** 2191–2198.

WEB SITE REFERENCES

http://www.ensembl.org; ENSEMBL Project home page.
http://genome.ucsc.edu; UCSC Genome Bioinformatics.
http://homepages.ed.ac.uk/eang33/mcinstructions.html; MCALIGN home page.
http://www1.imim.es/~castresa/Gblocks/Gblocks.html; GBLOCKS home page.
http://ncbi.nlm.nih.gov/dbEST; NCBI EST database home page.
ftp://ncbi.nlm.nih.gov/repository/UniGene; NCBI UniGene FTP site.
ftp://ncbi.nlm.nih.gov/pub/sage; NCBI SAGE FTP site.
ftp://ncbi.nlm.nih.gov/refseq; NCBI RefSeq FTP site.

Received May 28, 2003; accepted in revised form February 27, 2004.