Variation in the Ratio of Nucleotide Substitution and Indel Rates across Genomes in Mammals and Bacteria

Jian-Qun Chen,*† Ying Wu,*† Haiwang Yang,* Joy Bergelson,*† Martin Kreitman,† and Dacheng Tian*

*State Key Laboratory of Pharmaceutical Biotechnology, Department of Biology, Nanjing University, Nanjing, China; and †Department of Ecology & Evolution, University of Chicago

Rates of nucleotide substitution and insertion/deletion (indel) are known to vary across the functional components of a genome. Little attention has been paid, however, to the quantitative relationship between the two. Here we investigate the ratio of nucleotide substitutions to indels (S/I) in different regions of 4 primates, 70 bacteria, and 8 other genomes. We find that the ratio differs at 5.4-times between coding and noncoding, 3.3-times between conserved and less conserved coding sequences, and 1.46-times between nonrepeat and repeat regions. The S/I ratio is also positively correlated with the level of divergence between the genomes compared. Our results suggest that the S/I ratio may reflect differences in the efficacy of selection against indels. Due to the sensitivity of indel density in different regions, this ratio varies over a much larger range. With the recent discovery suggesting that indels act as local enhancers of mutation in surrounding sequences, nucleotide substitution rates are expected to be accelerated in regions of low constraint, where indels tend to accumulate, but will otherwise be modulated in proportion to the level of a sequence’s functional constraint. Indels, therefore, may play a nontrivial role in controlling differences in genetic variation and divergence across functional regions of a genome.

Introduction

The high mutation rate of insertions and deletions (indels) and their contributions to genome and adaptive evolution have only recently begun to be explored (Denver et al. 2004). Single-nucleotide polymorphism (SNP) and evolution, in contrast, have been heavily studied for their role in nearly all biological processes, including genetic disease, gene evolution, adaptation, and species divergence. Indels are now known to play an important role in these same processes. Indels have been found to dramatically reduce the occurrence of crossovers on their surrounding regions (Hammarlund et al. 2005). Suppressed recombination around indels, it has been argued, can trigger a “diversification front,” which under certain circumstances can eventually lead to global sequence divergence (Vetsigian and Goldenfeld 2005). Insertions have also been revealed to radically increase ectopic recombination frequency (Sun et al. 2008), which will lead to chromosome instability and genetic variation. In addition, indels are mutational components of gene and pseudogene evolution (Nedelcu 2001; Zhang and Gerstein 2003) and are important in the long-term evolution of genome size (Petrov et al. 2000; Gregory 2004). Indels are often deleterious, as evidenced by their frequent association with human disease (Zoghbi and Orr 2000; Kondrashov and Rogozin 2004). But they can also be beneficial as, for example, frameshift mutational activation of genes needed for microbes to survive in changing environments (van Belkum et al. 1998; van der Woude and Baumler 2004). In addition, indels have been associated with apoptosis mechanisms (Ferro et al. 2001), phenotypic change (Burch et al. 1997), and activity of surface antigen (Rocha and Blanchard 2002). A striking pattern of covariation between SNPs and indels is known to occur both within and between species (Hardison et al. 2003; Zhang et al. 2008), indicating a common causal mechanism regulates the two types of mutation. Britten et al. (2003) found that indels dominate the process of early divergence, revealed by comparing the ratio of nucleotide substitutions to indels (S/I ratio) between closely related taxa. The S/I ratio generally increases as interspecies divergence increases, perhaps as a consequence of both the lack of correction for multiple overlapping indel events in more highly diverged sequences as well as possible selective constraints on local gene sequence length. Closely related sequence comparisons, therefore, more closely reflect the actual mutation rate of indels. Importantly, Denver et al. (2004) showed that there were more indels than base substitutions in the early divergence for Caenorhabditis elegans.

For obvious reasons, the S/I ratio is expected to be governed primarily by functional constraint of the sequences in which they occur. For example, a one- or two-nucleotide indel in a coding sequence results in a frameshift mutation that can abolish the activity of the gene product (Garcia-Diaz and Kunkel 2006), whereas most “synonymous” mutations will be at best only weakly selected. Indels with non-3n size multiples will be subjected to strong purifying selection in coding regions (Taylor et al. 2004), and a higher S/I ratio will be expected.

In addition to their direct role in gene and genome evolution, we have recently shown that nucleotide substitution rates increase inversely with distance from indels over a range of a couple of hundred bases and that this increase is in part a consequence of an increase in the local mutation rate (Tian et al. 2008). In yeast, we have estimated this indel-induced mutation rate to be 10-fold higher than the underlying mutation rate.

With the possibility of indel-induced mutation in mind, we were interested in investigating the relationship between indel occurrence and functional constraint on the sequences in which they occur. The rationale for this interest is that indel density may be more strongly governed by natural selection than nucleotide mutation in relation to the functional constraint on a sequence. If this is the case, then indels can have an important role in governing...
substitution rate differences across functional regions of the genome. A comprehensive study on S/I ratio could shed light in the role played by indels. Here we systematically investigated the S/I ratio by comparing pairs of aligned genome sequences from the same species (human, bacteria, two rice, and three yeast) and from closely related species (three primates, mouse, and rat). The analysis of the S/I ratio across these genome comparisons allows us to explore how this ratio changes across functional regions of the genome and between distinctly different taxonomic groups.

Materials and Methods

Genome Sequences

Four mammalian genome sequences of three species were selected (table 1). Two human (Homo sapiens) genomes were from International Human Genome Sequencing Consortium (HGSC, released Build 36, which contains 2,850 Mb nucleotides) and from Celera Homo sapiens whole genome (ftp://ftp.ncbi.nlm.nih.gov/). Chimpanzee (Pan troglodytes) genome was from the Chimpanzee Sequencing and Analysis Consortium (released Build 2 version 1), and Rhesus (Macaca mulatta) genome was produced by the Baylor College of Medicine Human Genome Sequencing Center.

The bacterial genome sequences were also selected to investigate the correlation between nucleotide substitutions and indels. The filtered species satisfied the criteria that genome sequences were available for at least two different strains of the same species, and the nucleotide diversity between two strains was <5%. In total, 25 species (70 qualified bacterial genomes) were analyzed. These genomic sequences, downloaded from GenBank (http://www.ncbi.nlm.nih.gov/genomes/), belong to nine groups: Actinobacteria, Bacteroidetes/Chlorobi, Betaproteobacteria, Chlamydiae/Verrucomicrobia, Deinococcus-Thermus, Firmicutes, Gammaproteobacteria, Spirochaetes, and Tenericutes.

The other eight genomes are mouse (Mus musculus) and rat (Rattus norvegicus), dog (Canis lupus familiaris), two rice lines (Oryza sativa L. var. Nipponbare vs. var. 93-11), and three baker’s yeast strains (Saccharomyces cerevisiae str. S288C vs. RM11-1a and S288C vs. YJM789). The primate sequences were from the University of California–San Cruz (UCSC) Genome Browser (http://genome.ucsc.edu/) and the others from GenBank.


Sequence Alignments

The choice of an alignment algorithm is crucial to obtaining a reliable S/I ratio because indels can be the primary source of alignment uncertainty. ClustalW is one of the most widely used methods to align multiple sequences. BlastZ, on the other hand, was designed to identify orthologous sequences in genome-wide comparisons (Schwartz et al. 2003). To compare the performance of these two algorithms on the S/I ratio, we aligned approximately 1.5 Mb of human sequences from the HGSC database with four contigs of human sequence from Celera database, eight of chimpanzee sequence, and eight of rhesus sequence by both BlastZ and ClustalW (manual correction was used in region of obviously unpaired sequences). The two alignment methods produced similar results, especially in the less divergent alignments (supplementary table S2, Supplementary Material online). As expected, the two alignment methods produced nearly identical results for the human/human alignment. In contrast, the alignment of human/rhesus (5.56% divergence) is 30% shorter by BlastZ than by ClustalW, and the S/I ratio is 18% less. Such a difference is expected because only clearly defined indels (the simple indels and indels ≤300 bp) are retained in BlastZ alignments.

On the whole, the S/I ratios produced by the BlastZ algorithm were highly correlated with those from ClustalW alignments (e.g., r = 0.949, 0.974, or 0.980 between the ratios from the two methods in the comparisons of human/rhesus, human/chimpanzee, or human/human, respectively, in Table 1: Sequence Statistics in the Comparisons among Primates and Bacteria

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Method/Region</th>
<th>Length (Mb)</th>
<th>Substitution</th>
<th>Indel</th>
<th>S/I</th>
<th>Divergence (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human/human</td>
<td>Genome</td>
<td>647</td>
<td>717,772</td>
<td>151,966</td>
<td>4.72</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>Coding</td>
<td>11</td>
<td>7,698</td>
<td>676</td>
<td>11.39</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>Noncoding</td>
<td>637</td>
<td>710,074</td>
<td>151,290</td>
<td>4.69</td>
<td>0.11</td>
</tr>
<tr>
<td>Human/chimpanzee</td>
<td>Genome</td>
<td>485</td>
<td>6,369,026</td>
<td>725,470</td>
<td>8.78</td>
<td>1.32</td>
</tr>
<tr>
<td></td>
<td>Coding</td>
<td>7</td>
<td>45,177</td>
<td>1,168</td>
<td>38.68</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>Noncoding</td>
<td>478</td>
<td>6,323,849</td>
<td>724,302</td>
<td>8.73</td>
<td>1.33</td>
</tr>
<tr>
<td>Human/rhesus</td>
<td>Genome</td>
<td>192</td>
<td>11,665,775</td>
<td>1,018,423</td>
<td>11.45</td>
<td>6.34</td>
</tr>
<tr>
<td></td>
<td>Coding</td>
<td>13</td>
<td>89,488</td>
<td>1,467</td>
<td>61.00</td>
<td>3.03</td>
</tr>
<tr>
<td></td>
<td>Noncoding</td>
<td>189</td>
<td>11,576,287</td>
<td>1,016,956</td>
<td>11.38</td>
<td>6.40</td>
</tr>
<tr>
<td>Total in primate</td>
<td>Genome</td>
<td>1,324</td>
<td>18,752,573</td>
<td>1,895,859</td>
<td>9.89</td>
<td>1.43</td>
</tr>
<tr>
<td></td>
<td>Coding</td>
<td>20</td>
<td>142,363</td>
<td>3,311</td>
<td>43.00</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>Noncoding</td>
<td>1,304</td>
<td>18,610,210</td>
<td>1,892,548</td>
<td>9.83</td>
<td>1.44</td>
</tr>
<tr>
<td>Total in bacteria</td>
<td>Genome</td>
<td>212</td>
<td>2,191,676</td>
<td>111,782</td>
<td>19.61</td>
<td>1.04</td>
</tr>
<tr>
<td></td>
<td>Coding</td>
<td>179</td>
<td>1,814,482</td>
<td>45,714</td>
<td>39.69</td>
<td>1.02</td>
</tr>
<tr>
<td></td>
<td>Noncoding</td>
<td>33</td>
<td>377,194</td>
<td>66,068</td>
<td>5.71</td>
<td>1.15</td>
</tr>
</tbody>
</table>

Note.—S/I represents the ratio of nucleotide substitutions to indels; the length denotes the total alignments.
supplementary table S2, Supplementary Material online) but tended to be more conservative. For these reasons, we chose the BlastZ as alignment method for all data analyses. The BlastZ scoring matrix was the same as the one UCSC used for the pairwise alignments of human and chimpanzee.

When there is an indel >300 bp or an ambiguous nucleotide in aligning, the aligned sequence was sectioned into two alignments. Therefore, the indel size ranges from 1 to 300 bp in the alignments used for our analysis. To assure the orthology of alignments used in the analysis, we removed all alignments that were <10 kb in length.

In the comparison of mammalian genomes, 10 human chromosomes (13–22) were selected to represent the whole genome. This set of human sequences was used to align with corresponding sequences from another human, chimpanzee, and rhesus.

Sequence Analysis

Perl programs were used to calculate the number of nucleotide substitutions and the number and length of indels throughout each alignment. The length of each indel was calculated according to the gap length in an alignment. More than 98.5% of indels are no more than 100 bp in both primates and bacteria, and their distribution is demonstrated in supplementary figure S1 (Supplementary Material online). Average nucleotide divergence (or diversity) was adjusted with the Jukes and Cantor correction for multiple hits (the corrected rate was not used for the calculation of S/I ratio, which was directly from the number of nucleotide substitutions). Instead of indel number, we also calculated the S/I ratio using the length difference between orthologous sequences in bacteria, considering that there is at least one indel in a protein-coding region or a spacer between genes with length difference no matter what aligning method is used. A consistent result was also obtained in coding regions (supplementary fig. S2, Supplementary Material online), which constitute most of bacterial genomes, indicating the feasibility of our methods.

Strand slippage may play an important role in the generation of indels in repetitive DNA sequences (Streisinger et al. 1966). Therefore, each indel nucleotide sequence was compared with the immediately adjacent 5’ and 3’ sequence, and indels with the identical (repetitive) adjacent words (base pairs) >3 bp or repeat units on either or both sides were categorized as being “slippage-like.” The indel sequence itself was excluded from the count. This method for characterizing indels is consistent with previous study of indel-producing mechanisms (Taylor et al. 2004). We investigated the accumulation of slippage-like indels over time, by analyzing the proportion of such indels in relation to total indels on chromosome 1 in the comparisons of mammal genomes.

Bacterial genes with the start and stop codon were selected to estimate the proportions of synonymous and “nonsynonymous” differences by the Nei and Gojobori (1986) method.

Results

Variation of the S/I Ratio among Genomes

We calculated S/I ratios for 4 primates and for 70 bacteria in whole-genome sequences (table 1; supplementary table S1, Supplementary Material online). Two results stand out. First, the average S/I ratio for bacteria (19.6) is about 2-fold larger than for primates (9.9) overall, and the S/I difference is much greater in coding regions than in noncoding regions for both bacteria (39.7 vs. 5.7) and primates (43.0 vs. 9.8). This difference is explored in more detail in the Variation of S/I Ratio in Coding Noncoding Regions.

Second, within both bacteria and mammals, the indel divergence rate increases more slowly than nucleotide divergence (supplementary table S3, Supplementary Material online and figs. 1 and 2). This trend is clearly seen in the set of three comparisons using human versus human, chimpanzee, and rhesus genome sequences (divergence estimated at ≈1, 7, and 25 My, respectively; Stewart and Disotell 1998; Wildman et al. 2003). These comparisons reveal an S/I ratio of 4.72, 8.78, and 11.45, respectively. When the nucleotide divergence (D) is used to measure the time of divergence, a curve can be obtained between D and S/I ratio (r = 0.886 and P < 0.001; fig. 1A). Similarly, there is a significant correlation between D and S/I ratio (r = 0.725 and P < 0.001) among bacteria genomes (fig. 1B), where it is seen to increase with even slight increases in the intraspecific divergence of the pairs of bacteria.

The correlation between the divergence time and S/I ratio among genomes suggests that there may be also a corresponding correlation between the divergence in a region and the ratio in this region. We categorized the alignments into small groups by the order of nucleotide divergence levels in each alignment (fig. 1C–H). A positive relationship is observed between two human genomes, human and chimpanzee, human and rhesus, mouse and rat, two rice lines (Nip and 9311), and three yeast strains (S288, Rml1, and YJM789), respectively. The significantly positive correlation in the comparisons among regions with different levels of divergence is consistent with the relationship in the comparisons between genomes.

Variation of S/I Ratio in Coding and Noncoding Regions

Indels are expected to be deleterious when they occur in functional sequences, especially coding regions, where they can induce frameshifts in the translation frame as well as adding to or subtracting from the length of the encoded protein. To test this prediction, we calculated S/I ratios in coding and noncoding regions separately in the comparisons between primates and between bacteria. As shown in figure 2A and B, the S/I ratio in coding sequences is substantially greater than in noncoding regions, and the ratio is more nearly constant over time in both coding and noncoding regions. The fact that S/I ratios are not substantially different in mammals and bacteria perhaps indicates a universal constancy in the relative occurrence of the two kinds of mutation. It has long been noted that nucleotide mutation rates are not substantially different between prokaryotes and eukaryotes (Ochman and Wilson 1987), which is consistent with our results.

All the primate and bacterial comparisons reveal a significant linear relationship between the S/I ratio and divergence (fig. 2). Linear regression analysis of S/I against
divergence (supplementary table S3, Supplementary Material online) indicates that primate and bacterial comparisons have similar slopes ($b$) and intercepts ($a$). Statistically, $b_1$ (primate-coding), $b_3$ (bacteria-coding), and $b_5$ (combined-coding) or $b_2$ (primate-noncoding), $b_4$ (bacteria-noncoding), and $b_6$ (combined-noncoding) are not significantly different (they are combined into $b_5$ and $b_6$, supplementary table S3, Supplementary Material online). By comparing the coding:noncoding slopes, it can be seen that the relative coding S/I ratio is 16.0 times ($b_5$/$b_6$).

Additionally, previous study has shown that bacterial noncoding regions are conserved relative to synonymous sites (Hughes and Friedman 2004). We therefore investigated the ratio of synonymous substitutions to indel ($sS/I$) and nonsynonymous substitutions to indel ($nS/I$) in protein-coding genes of bacteria. As expected, both $sS/I$ and $nS/I$ ratio are highly correlated with divergence ($P < 0.01$ for both), and as shown in supplementary figure S3 (Supplementary Material online), the slope of nonsynonymous sites ($h_n = 1107$) is significantly greater than that of synonymous sites ($h_s = 518$). The rate of synonymous substitutions per site (0.0464) is higher than the divergence of noncoding regions (0.0115), whereas the rate of nonsynonymous substitutions is extremely low (0.0047). The
conserved noncoding regions suggest that $b_4$ should be higher than $b_s$. However, the real $b_s$ cannot be calculated because there is no synonymous indels in coding regions. Therefore, the $S/I$ ratio, calculated by using nonsynonymous indels, may not be comparable to $S/I$ ratio in noncoding regions.

Variation of Substitutions/Indels in Conserved and Less Conserved Coding Regions

The large difference in $S/I$ ratios in coding and noncoding sequences suggests that more functionally constrained genes, as measured by coding substitution rates, will have higher $S/I$ ratios than less constrained genes. The more conserved genes have fewer nucleotide substitutions to estimate rates, so we analyzed all the gene pairs on chromosomes 13–22 in the comparisons of human–rhesus and human–chimpanzee. In total, there are 844 and 2,708 whole gene pairs in aligned sequences. The genes were ordered by levels of coding region nucleotide divergence and pooled into three divergence intervals (0–0.03, 0.03–0.05, >0.05 for the comparison of human–rhesus and 0–0.007, 0.007–0.015, >0.015 for the comparison of human–chimpanzee). As expected, the $S/I$ ratio is positively related to gene conservation (i.e., negatively correlated with $D$) in the human–rhesus and human–chimpanzee comparisons (fig. 2C and D).

Substitutions/Indels in Repeat and Nonrepeat Regions

The predominant mutational mechanism creating indels is thought to be polymerase slippage, also known as slipped-strand mispairing (Levinson and Gutman 1987; Li et al. 2002). The mispairing of DNA strands during replication or recombination can result in a single-stranded loop that may cause an indel mutation (Garcia-Diaz and Kunkel 2006). The propensity for polymerase slippage is positively correlated with repeat length (Klintschar and Wiegand 2003); infidelity in repeat sequence pairing is likely to be the major reason for slippage-like indel formation. Regions with repetitive sequences exhibit a higher-than-background frequency of indel (Nishizawa M and Nishizawa K 2002). Therefore, a higher density of indels is expected in repetitive sequences, which will decrease the $S/I$ ratio. However, there may also be a greater potential for underestimating the indels events if they tend to be overlapping or indistinguishable from one another.

To investigate the effects of repeat sequences on the $S/I$ ratio, we calculated the ratios separately for repeat and nonrepeat regions (by RepeatMarsk) in the comparisons of human–human, human–chimpanzee, and human–rhesus in chromosomes 13–22 (fig. 3A and B). Clearly, the ratio in nonrepeats is significantly higher than the one in repeat regions in the comparisons of human–human (5.95 vs. 4.08) and human–chimpanzee (9.85 vs. 7.93). The difference in the ratios is not significant in the comparison between human and rhesus (11.39 vs. 11.53), however, perhaps reflecting an underscoring of repeat indels in this more distantly related comparison. Thus, once again, the $S/I$ ratio increases with divergence for both nonrepeat and repeat and that the higher indel occurrence can only be observed in the less divergent comparisons. Indeed, the relative increase (2.10) of indels ($I_{\text{repeat}}/I_{\text{nonrepeat}}$, the average indels per kilobase in repeat regions divided by those in nonrepeats) is far greater than the relative increase ($1.44, D_{\text{repeat}}/D_{\text{nonrepeat}}$) of $D$ in repeat over nonrepeat regions in the comparison of human–human, indicating that the indel rate varies more than the rate of substitutions does from region to region (table 2; fig. 3C). However, the relative indel increase is much less (1.44 vs. 1.16) in the human–chimpanzee comparison and disappears in the human–rhesus comparison (1.16 vs. 1.17).

To further examine the potential undercounting of indels in repetitive regions, especially in the more diverged comparisons, we calculated the proportion of slippage-like indels in...
As expected, the proportion of slippage-like indels (see Materials and Method for definition) is negatively correlated with the process of divergence in the primate comparisons (Fig. 3D). For example, when the same chromosome 1 was compared, the proportion of such indels is highest between human and human (37.16%) and declines as the divergence time increases (human vs. chimpanzee > human vs. rhesus > mouse vs. rat > dog vs. mouse > human vs. dog > human vs. mouse or rat). The proportion is the lowest in the comparison of human versus rat (7.32%). Further calculation of these proportions revealed that the decline of slippage-like indels occurred basically in repeat sequences. Figure 3E showed that the decline in repeat regions is from 45.43% between human individuals to 19.99% between human and rhesus. In contrast, the decay in nonrepeat sequences is only from 25.05% to 18.87% in the corresponding comparisons.

Interestingly, a negative correlation is not observed in the bacteria genome comparisons (fig. 3F), probably the result of having a smaller proportion of noncoding regions and fewer repeat sequences in their genomes.

Discussion

The S/I ratio is a convenient gage of nucleotide substitution propensity scaled to the indel rate. A major drawback of this measure is the well-accepted fact that nucleotide substitution rates are not constant over evolutionary time or across the genome. Nevertheless, the comparison of these ratios has revealed large differences that we believe are more parsimonious with rate variation in indels rather than with nucleotide substitution. Below we explore factors contributing to this indel rate variation. We complete the discussion...
by considering effects indel density has on substitution rate heterogeneity, a consequence of mutation induction surrounding segregating indels.

Mechanisms Underlying Variation in the S/I Ratio

First, we found in both primate and bacterial sequence comparisons, the rate of substitution indel increases in relation to sequence divergence. This increase in the substitution indel rate is unlikely to have resulted from a similar change in rates over time in both taxonomic groups, and so we attribute the increase in the S/I ratio with divergence time primarily to systematic underestimation of indels in more distantly related sequences. Such a finding is not unexpected as the algorithms for aligning sequences generally impose gap opening penalties for introducing an indel that tend to minimize (or at least decrease) the apparent number of indels in aligned sequences. Our analysis of indel rates showing that the S/I ratio declines faster in repetitive than nonrepetitive regions, where overlapping indels are more likely, reinforces this argument. In the absence of a provably good algorithm for correcting the indel number to account for multiple (i.e., overlapping) events, perhaps the S/I ratio can be used to fine-tune sequence alignment algorithms to produce a more accurate representation of the indel rates across divergence time (i.e., produce a more nearly constant S/I ratio). In any case, our study confirms previous studies showing that the indel rate is high—the same order of magnitude as nucleotide substitutions—and is best estimated between very closely related taxa. This is the reason our study focused on pairs of closely related taxa to estimate and compare S/I ratios.

As expected, we found indels to be less abundant in coding regions than in noncoding regions. More interestingly, we found that indel density decreases faster than nucleotide substitution rates when comparing coding regions under different levels of selective constraint. To account for these observations, we hypothesize that, on average, indels are more deleterious than nucleotide substitutions (including nonsynonymous changes). As with nucleotide substitutions, they will also have a broad distribution of selective effects, ranging from completely neutral to lethal. We also hypothesize that, in general, selective effects of indels and nonsynonymous mutations in the same codons or functional regions of a protein are correlated but with indels increasing in their negative selective effects faster than amino acid replacement changes. Thus, as selective constraint on protein function increases, the rate of both amino acid replacement and indel evolution will decrease but with the indel rate declining faster.

The comparison of coding and noncoding S/I ratios between primate and bacterial sequences provides additional information about the evolutionary forces acting on indels. First, as shown in figure 2A and B, the rates of noncoding S/I are very nearly indistinguishable between bacterial and primate comparisons. It is by no means certain that bacterial and mammalian noncoding regions have similar overall densities of functionally constrained sequences (such as cis-regulatory motifs). Nevertheless, on the assumption, this is the case and further assuming most noncoding indel substitutions are selectively neutral, the equivalent rates of noncoding S/I between bacteria and mammals suggest similar underlying origination rates (mutation) for indels in the two animal kingdoms. This perhaps suggests that the mechanisms in DNA replication creating indel mutations are universal.

Comparing the S/I ratio between coding and noncoding or between conserved and less conserved sequences, we observed a significant difference in the S/I ratios, for example, 5.4-fold (coding vs. noncoding) and 3.3-fold difference (conserved vs. less conserved) in the comparison of human versus rhesus, respectively. Such a large difference gives a clear indication that indels are eliminated at a faster rate in coding regions.

In addition, there may be constraints on the magnitude of length change accepted by the evolutionary process. Such constraints are almost certainly imposed by spacing requirements between promoters and coding regions, between regulatory factor binding sites, and other chromatin structural constraints. These constraints create selective pressures against progressive changes in the length of a sequence interval, thus imposing a form of stabilizing selection on indel mutation. Such a process is expected to further contribute to underestimation of independent indel mutations.

Table 2
The Relative Variation of Nucleotide Divergence or Indel Density from Region to Region in the Comparisons among Primates

<table>
<thead>
<tr>
<th>Coding versus Noncoding</th>
<th>D_{coding} (%)</th>
<th>D_{noncoding} (%)</th>
<th>D_{n–c} (%)</th>
<th>D_{n–c} (%)</th>
<th>I_{coding} (%)</th>
<th>I_{noncoding} (%)</th>
<th>I_{n–c} (%)</th>
<th>I_{n–c} (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human/human</td>
<td>0.074</td>
<td>0.112</td>
<td>0.038</td>
<td>1.517</td>
<td>0.006</td>
<td>0.024</td>
<td>0.017</td>
<td>3.680</td>
</tr>
<tr>
<td>Human/chimpanzee</td>
<td>0.670</td>
<td>1.322</td>
<td>0.652</td>
<td>1.973</td>
<td>0.017</td>
<td>0.151</td>
<td>0.134</td>
<td>8.739</td>
</tr>
<tr>
<td>Human/rhesus</td>
<td>2.966</td>
<td>6.130</td>
<td>3.164</td>
<td>2.067</td>
<td>0.049</td>
<td>0.539</td>
<td>0.490</td>
<td>11.076</td>
</tr>
<tr>
<td>Repeat versus nonrepeat</td>
<td>D_{repeat} (%)</td>
<td>D_{nonrepeat} (%)</td>
<td>D_{r–n} (%)</td>
<td>D_{r–n} (%)</td>
<td>I_{repeat} (%)</td>
<td>I_{nonrepeat} (%)</td>
<td>I_{r–n} (%)</td>
<td>I_{r–n} (%)</td>
</tr>
<tr>
<td>Human/human</td>
<td>0.132</td>
<td>0.092</td>
<td>0.040</td>
<td>1.441</td>
<td>0.032</td>
<td>0.015</td>
<td>0.017</td>
<td>2.10</td>
</tr>
<tr>
<td>Human/chimpanzee</td>
<td>1.419</td>
<td>1.221</td>
<td>0.198</td>
<td>1.162</td>
<td>0.179</td>
<td>0.124</td>
<td>0.105</td>
<td>1.44</td>
</tr>
<tr>
<td>Human/rhesus</td>
<td>6.660</td>
<td>5.694</td>
<td>0.966</td>
<td>1.170</td>
<td>0.577</td>
<td>0.050</td>
<td>0.078</td>
<td>1.16</td>
</tr>
<tr>
<td>Conserved versus less conserved</td>
<td>D_{conserved} (%)</td>
<td>D_{less conserved} (%)</td>
<td>D_{c–l} (%)</td>
<td>D_{c–l} (%)</td>
<td>I_{conserved} (%)</td>
<td>I_{less conserved} (%)</td>
<td>I_{c–l} (%)</td>
<td>I_{c–l} (%)</td>
</tr>
<tr>
<td>Human/human</td>
<td>0.396</td>
<td>2.233</td>
<td>1.837</td>
<td>5.633</td>
<td>0.009</td>
<td>0.086</td>
<td>0.076</td>
<td>9.098</td>
</tr>
<tr>
<td>Human/rhesus</td>
<td>2.099</td>
<td>7.134</td>
<td>5.125</td>
<td>3.551</td>
<td>0.026</td>
<td>0.292</td>
<td>0.266</td>
<td>11.276</td>
</tr>
</tbody>
</table>

Note.—D_{coding} (D_c), D_{noncoding} (D_n), D_{repeat} (D_r), D_{nonrepeat} (D_nr), D_{conserved} (D_c), and D_{less conserved} (D_l) represent the divergence (D) in coding, noncoding, repeat, nonrepeat, conserved, and less conserved regions, respectively. Similarly, I_{coding} (I_c), I_{noncoding} (I_n), I_{repeat} (I_r), and I_{nonrepeat} (I_nr) represent indel density in corresponding regions. The conserved genes are those with D < 0.007 between human/chimpanzee (1,363 genes) or D < 0.03 between human/rhesus (321 genes), and the less conserved are genes with D > 0.015 between human/chimpanzee (452 genes) or D > 0.05 between human/rhesus (250 genes).
We found a nonrandom occurrence of indels in repeat regions. By our strict definition of slippage-like indels, a substantial fraction of indels (45.43%) in the repeat regions of human–human alignments were found to have a 4-bp (units) sequence identical to the inserted or deleted sequence, directly adjacent to the indel event (fig. 3E). This was significantly more than expected by chance and was not simply a consequence of expansion and contraction of repetitive sequences (Taylor et al. 2004). Similar findings have been reported by Nishizawa M and Nishizawa K (2002). The prevalence of slippage-like indels is believed to be a consequence of mutational mechanism rather than selection (Taylor et al. 2004). The S/I ratio is strongly influenced by the nonrandom occurrence of slippage-like indels as shown in figure 3B. The human–human comparison displayed an S/I ratio 1.46 (5.95/4.08) times higher in nonrepeat versus repeat sequences. This is certainly an underestimate as the amount of reciprocal or overlapping indel events must be more common in repetitive regions.

Influence of Indel Density on Nucleotide Substitution Rate Heterogeneity across Genomes

Having shown here that the rate of substitution indel varies across genomes and is positively correlated with functional constraint, but indel is more sensitive to it than nucleotide substitution. We complete the discussion by attempting to tie this observation with another recent discovery we have made about the mutagenic effect of segregating indels on surrounding DNA. In particular, Tian et al. (2008) showed that nucleotide substitution rates are elevated surrounding indels in comparisons of genome sequences across many different taxonomic groups. A closer analysis of yeast genome data, comparing nucleotide substitution rates in the DNA surrounding indel mutations (derived alleles) with alleles containing the ancestral state (no indel), indicated approximately a 10-fold relative increase in the nucleotide mutation rate within 100 bp of an indel in indel heterozygotes.

Weak mutagenic activity of segregating indels and large differences in the density of indels across the genome in relation to functional constraint suggests to us that nucleotide substitution rates will tend to be highest in regions of lowest functional constraint and lowest in the most functionally constrained regions. Indel density, therefore, has the consequence—perhaps fortuitously—of modulating nucleotide mutation rates in an adaptive direction, reducing the mutation rate in regions where they are most likely to be harmful. Of course, repetitive DNA, a major factor in indel formation, also is highest in regions of the genome deemed to be lowest in functional constrained regions is strong enough to regulate the repetitiveness of DNA, and thus the indel rate or, conversely, whether repetitiveness is selected to be higher in certain regions, thus promoting mutation via higher indel rates, deserves future consideration. In this regard, Comer PN, Kreitman M. 2002. Population, evolutionary and genomic consequences of interference selection. Genetics. 161:389–410.


Douglas Crawford, Associate Editor

Accepted March 24, 2009