Majority of divergence between closely related DNA samples is due to indels

Roy J. Britten*†, Lee Rowen‡, John Williams*, and R. Andrew Cameron*

*California Institute of Technology, 101 Dahlia Avenue, Corona del Mar, CA 92625; and ‡Institute for Systems Biology, 1441 North 34th Street, Seattle, WA 98103

Contributed by Roy J. Britten, February 15, 2003

It was recently shown that indels are responsible for more than twice as many unmatched nucleotides as are base substitutions between samples of chimpanzee and human DNA. A larger sample has now been examined and the result is similar. The number of indels is ≈1/12th of the number of base substitutions and the average length of the indels is 36 nt, including indels up to 10 kb. The ratio (R_u) of unpaired nucleotides attributable to indels to those attributable to substitutions is 3.0 for this 2 million-nt chimp DNA sample compared with human. There is similar evidence of a large value of R_u for sea urchins from the polymorphism of a sample of *Strongylocentrotus purpuratus* DNA (R_u = 3–4). Other work indicates that similarly, per nucleotide affected, large differences are seen for indels in the DNA polymorphism of the plant *Arabidopsis thaliana* (R_u = 51). For the insect *Drosophila melanogaster* a high value of R_u (4.5) has been determined. For the nematode *Caenorhabditis elegans* the polymorphism data are incomplete but high values of R_u are likely. Comparison of two strains of *Escherichia coli* O157:H7 shows a preponderance of indels. Because these six examples are from very distant systematic groups the implication is that in general, for alignments of closely related DNA, indels are responsible for many more unmatched nucleotides than are base substitutions. Human genetic evidence suggests that indels are a major source of gene defects, indicating that indels are a significant source of evolutionary change.

Mutations in the DNA are the source of variation in Darwinian evolution. Therefore it is likely that the examination of DNA differences between closely related species or among polymorphic variations in DNA of a given species will give insight into the nature of the mutations and the process of evolution. In the present paper, published and unpublished data are summarized for examples from several distantly related phylogenetic groups, and the data show that indels dominate the process of early divergence. There is a continuing problem in these data of the upper limit in the size of detected gaps and bias against larger ones. The groups sampled are apes (chimp–human DNA comparison), sea urchins (*Strongylocentrotus purpuratus* polymorphism), bacteria (*Escherichia coli* substrain comparison), insects (*Drosophila* polymorphism), nematodes (*Caenorhabditis elegans* polymorphism), and plants (*Arabidopsis* polymorphism). It is also noted that human genetic diseases are frequently caused by indels. The first part of the paper summarizes the results for samples of chimp DNA compared with the human genome sequence. Then an example of sea urchin polymorphism is briefly described. Initial comparison of two strains of *E. coli* O157:H7 is described. Finally, the published polymorphism data are reviewed and brought together with the data reported here to draw the conclusion that indel formation is a major and significant evolutionary process.

Materials and Methods

Chimpanzee bacterial artificial chromosome (BAC) sequences have been listed in GenBank by two groups (Genome Center, University of Oklahoma, Norman, and the Human Genome Sequencing Center, Baylor College of Medicine, Houston) and have been downloaded. The attempt has been made in every case to align the complete chimp BAC sequence with the human genome, regardless of the presence of repeated sequences, which typically consist of about half of the BAC sequence. The repeated sequences, naturally, sometimes complicate the alignment process. The National Institutes of Health program "BLAST the Human Genome" was used to find the most promising region of the human genome for alignment with each particular chimp BAC sequence. This program works well because the human repetitive sequences are filtered out during the comparisons and then apparently reinserted for mapping the results. Usually only one region of the human genome shows a full or nearly full alignment with a chimp BAC sequence, whereas other regions show short or fragmentary alignments. Where duplications of long regions have occurred as on chromosome 22 there is uncertainty and we have not included these comparisons. For the next stage in the analysis a program has been written that almost always accurately detects mismatches and gaps in the alignment. It is called GAPD for gap detection or gap determination and is described in the next few sentences. Standard sequence comparison programs such as Smith Waterman are used to find the human sequence that aligns with the start of the chimp BAC. From this aligned start GAPD goes nucleotide by nucleotide checking for mismatches. If a mismatch is seen, then a check is made of the succeeding 10 nucleotides, and if at least 6 of these match, the original mismatch is taken to be due to a base substitution. There is a possibility that there is a local region with many mismatches, and so the program looks successively at four 10-nt regions searching for a good match (6 of 10). If none are found it is presumed that a gap is present and possible registration or phase differences between the two sequences are tested, looking for 20-nt regions that match well (16/20). Usually a small registration difference suffices because most gaps are 1 or a few nucleotides. If not, larger registration differences are tested up to a limit of 10,000 nt. The longest gap certainly observed by this method is ≈9.5 kb. There are longer gaps, but this program is restricted to the gaps <10 kb. When a gap is found the program continues with the new registration until another mismatch is recorded or a gap is found, and so on to the end of the BAC. This is a specific program that works well with very similar sequences such as those of chimp and human DNA. It is the same program as used for the previous chimp human comparison (1), but some parameters have been changed. This method has been checked as described in succeeding paragraphs.

Experience has shown that rarely a region of extensive mismatch is followed by two successive good copies of a repeated sequence. In this case the registration difference the program proposes can be in error, and as soon as the repeated sequence ends the sequences no longer match. Then the program searches and finds the correct registration, recording an artificial gap in the other sequence to compensate for the first erroneous gap. These pairs of proposed gaps are manually recognized or identified by the following check.

Abbreviation: BAC, bacterial artificial chromosome.

†To whom correspondence should be addressed. E-mail: rbritten@caltech.edu.

EVOLUTION

Table 1. Characteristics of chimp BAC alignments

	Indels			Substitutions			
Name	No.	Total, nt	Avg. length, nt	No.	% in CpG	Length, nt	R_u
AC123983	154	4,622	30	2,265	24.4	131,422	2.0
AC123982	324	14,466	45	3,328	23.6	175,062	4.3
AC125393	206	2,491	12	2,584	18.4	140,578	1.0
AC125391	175	5,754	32	2,126	25.8	147,977	2.7
AC096630	173	3,841	22	2,867	23.3	160,599	1.3
AC093572	215	8,517	40	2,649	30.2	195,648	3.2
AC097335	225	5,467	24	2,683	19.0	139,478	2.0
AC007214	160	5,472	34	2,023	20.7	139,331	2.7
AC006582	190	4,068	21	1,905	20.7	100,217	2.1
AC118585	210	10,391	49	2,705	23.4	104,571	3.8
AC097265	259	15,640	60	2,930	29.0	167,612	5.3
AC120838	241	2,515	10	2,377	26.3	145,921	1.1
AC120782	321	20,652	64	3,924	27.5	160,136	5.3
AC124148	241	8,828	36	2,965	24.0	178,778	3.0

Length of sample, nt	2,087,300
Indel number	3,094
Indel length	112,728
Average indel length, nt	36.4
Base substitution count	37,331
R_u for the total sample	3.0

Column 1 is GenBank name. Column 2 is number of indels seen, column 3 is total length of indels, and column 4 is average length. Column 5 is number of substitutions and column 6 is percent substitutions occurring in CpGs. Column 7 is the length of the alignment. Column 8 is the ratio of indel total length to number of substitutions.

As a check each gap >100 nt is tested by using as a probe the sequence in the other species DNA opposite the gap. This sequence is compared with the complete chimp BAC and matching human sequence by using the Smith Waterman (slow) alignment program. If the gap is false, then there will be an accurately matching region in both species at the correct location, i.e., no gap. If the gap is real the match will be missing in the DNA of the correct species at the right location.

As a further check Smith Waterman alignments are made between the two DNAs by using segments of the size that the program can handle. Most gaps are confirmed, but occasionally this alignment program finds false alignments in the region of a gap consisting of many poorly matching short regions separated by short gaps, and these are ignored. The end result is a fairly precise list of gaps in the alignment of the chimp BAC with human DNA.

For comparison of the 5.5-megabase *E. coli* DNA sequences the first step was to cut them into 100-nt segments and compare each of the segments with all of the segments of the DNA of the other substrain by using BLAST. A plot of the difference in position in the two sequences of the most precisely matching pairs was very revealing. There were numerous short regions matched because of repeated sequences, but these could be ignored and the main alignment directly followed. The graph, which will be published elsewhere, exhibited the results of major rearrangements that had occurred. The regions demarcated by the large events showed many smaller indels. These regions were separated out and individually analyzed with GAPD.

Results

Chimpanzee–Human Comparisons. There are 25 chimp BAC sequences that we have studied. Of these, 14 can be aligned nearly from end to end with regions of the human genome, showing typical sequence divergence of 1–2% attributable to base substitutions. The remaining 11 BAC sequences are not easily aligned for their whole length. In two cases, at least, the difficulty in alignment was apparently due to the presence of very large indels, and these are discussed later. The focus has been on the >2 million nucleotides of successful long alignments. There are 3,094 gaps or apparent indels, and 93 of these are >100 nt long. About 57% of the length of these longer gaps corresponds to repeated sequences as judged from the sequence in the other species opposite the gap. The surrounding regions contain both repeated and single-copy sequences. The gaps occur about equally in both species DNA.

One direct observation shows that the bulk of mismatches are not due to sequencing errors. The fraction of the observed mismatches at CpGs ranges from 18% to 29% of all mismatches among the BAC alignments as shown in Table 1. Because CpGs occur at only about 1% or 2% of nucleotides in the human and chimp sequences this cannot have happened by chance. The fraction of CpGs that show differences between human and chimp DNA ranges from 13% to 20%. The reason is the well known biochemical mechanism that leads to the mutation of CpGs. The rate of mutation per nucleotide in CpGs is about 10-fold greater than for typical nucleotides in human DNA, as has been previously observed in *Alu* sequences (2). In a large sample of BAC end sequence comparisons (3), 15% of all CpG sites experienced changes between chimp and human compared with 1.24% average sequence difference attributable to base substitutions in this presumably single-copy DNA.

This evidence shows that the sequences are reasonably precise but does not prove that the gaps observed are true sequence differences rather than errors in assembly. The large number of gaps in alignment between DNAs of closely related species has not been observed previously, and ultimately should be independently confirmed as a natural process of mutation that occurs at the apparently observed rate. It is unlikely that any significant fraction of the gaps are due to assembly or alignment errors, and for the purposes of this paper we assume that they are all real. Table 1 shows the variation among the 14 BAC samples. The R_u is listed and varies from 1.0 to 5.3. The indels consist of 5.4% of

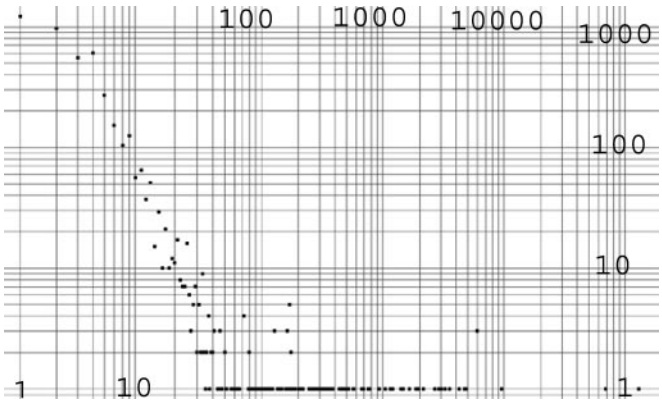

Fig. 1. The raw data on gaps between chimp and human alignments. Shown is log–log plot of number of gaps of a given size as a function of size. The vertical axis is the number of gaps and the horizontal axis is the gap length in nucleotides. The line near the bottom is all of the larger gaps, which are present only once with a given length. Gaps >5 kb are uncertain.

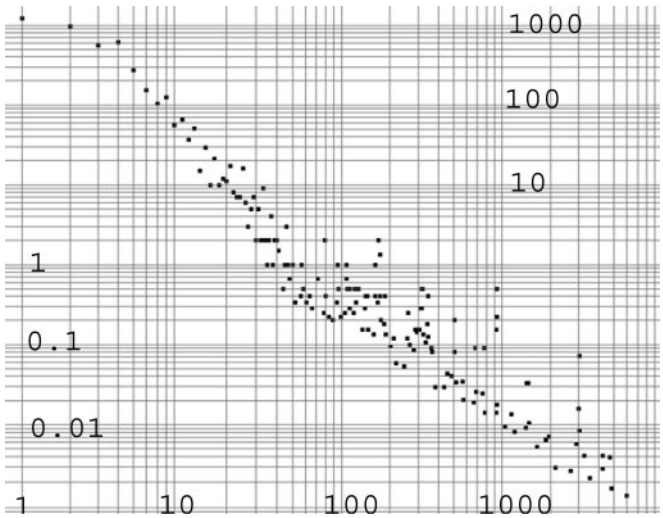

Fig. 2. The density of gaps vs. gap size. Shown is a log–log plot of the density function D_k against gap size. The horizontal axis is gap length in nucleotides. The vertical axis is the density function, which is the number of gaps of a given size divided by the spacing in length between gaps, which is the average of the difference in length to the next smaller gap and the difference in length to the next larger gap. Shown are gaps <5 kb.

the length of the DNA in this sample for gaps up to 10 kb long. The ratio (R_u) of unpaired nucleotides attributable to indels to those attributable to substitutions is 3.0 for this 2 million-nt chimp DNA sample compared with human. This number is for gaps <10 kb because the larger gaps are uncertain.

Of course a number of gaps in alignment between chimp and human DNA have been observed in the past involving *Alu* repeat insertions, instability of tandem arrays, and retroviral indels. Thus there is nothing novel about the process, only that its magnitude and generality have not been fully realized. Assuming the last common ancestor was 6 million years ago, the observed rate of change per nucleotide affected by indels is 4.7×10^{-9} per year in each lineage. The average size of indels is 36 nt in our current sample, and thus the rate of occurrence per event of insertion or deletion can be estimated as $\approx 1.2 \times 10^{-10}$ per year in each lineage. In comparison, the occurrence of base substitutions in our sample is 1.78% or 1.5×10^{-9} per year in each lineage. This is larger that that observed for single-copy DNA of 1.24% (3). The reason for this difference is presumably that our sample includes repeated sequences, some of which, such as the *Alu* repeats, are known to have a higher rate of base substitution.

The observations are graphed in Fig. 1, which is a log–log plot of the number of indels of a given size against the size of indels. There are many more smaller ones, and all of the largest gaps occur only once, which is responsible for the horizontal line at the bottom of the figure. In an attempt to make a quantitative model of the occurrence of gaps Gu and Li (4) have plotted the number of gaps per length defined as N_k for a variety of gene regions, comparing rodents vs. humans. They derive the equation $N_k = Ck^{-b}$, where k is the length of the indel and C and b are constants. Because many of the gap sizes observed here occur only once we have defined another parameter, which reduces to N_k for the smaller gaps, called D_k for the density of gaps as a function of length. D_k is the number of gaps of a given size divided by the spacing between gap sizes averaged for the spacing to the next smaller and next larger gaps. Fig. 2 shows a log–log plot of D_k vs. k for the chimp–human gap data. In the region between gap lengths of 5 and 100 the slope of these data suggests a value of b in the range observed by Gu and Li of ≈1.7–1.9 for indels. However, for smaller indels the formula does not apply to these data and for large indels the value of b has fallen distinctly.

A major interest is the total length of the indels in a given DNA sample. It matters what the contribution of the various sizes of indels is to the total length, and Fig. 3 represents the data for this purpose. Fig. 3 is a log–log plot of the cumulated length vs. size. The first point at the lower left is the 1,223 single-nucleotide indels. The second point adds the 970 2-nt indels to reach 3,163 total, etc. The curve begins to level near 100-nt indels and then shows the major contributions of the larger indels. In the region of the graph up to ≈5,000 nt the slope upward is striking. The data for large gaps beyond 5 kb are very sparse, and this last part of the curve will surely rise with more data.

Alignments with large indels have difficulties, but one example of 9,500 nt in AC118585 has been thoroughly checked. In this case, despite its size the Smith Waterman program agrees exactly with GAPD. Recently we have observed that AC123981 apparently has a 128-kb gap in the chimp sequence. The human region opposite the gap contains a normal mix of many kinds of repeated sequences and matches chromosome 14. Within this region are coded three genes and part of a fourth all identified by ESTs and mRNA data. However, there is a duplicate human region on chromosome 22 that includes the same sequence mostly to 99% accuracy. There are divergent local regions and in the official interpretation additional genes are present on the

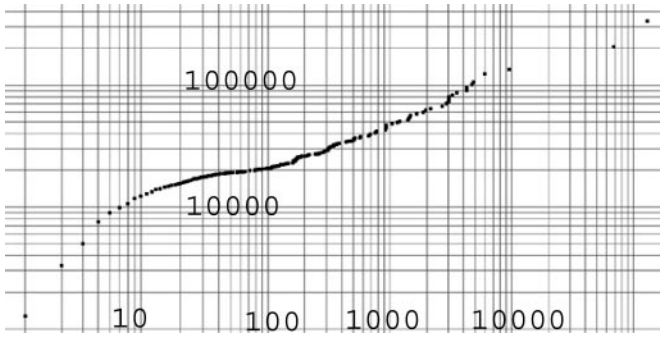

Fig. 3. The cumulative total of the length of gaps vs. gap size. The number of gaps of a given size is multiplied by the length of the gap and added to the previous total to obtain the cumulative total. The horizontal logarithmic axis is the gap size and the vertical logarithmic axis is the cumulative total. It is clear that the larger gaps contribute heavily. The last four points represent sparse data because long gaps are difficult to measure. New data could easily raise this part of the curve.

chromosome 22 segment, and none of the genes are identical to those identified on the chromosome 14 "copy." Therefore, this region is a good candidate for the identification of gene differences between chimp and human, but a careful analysis will have to wait for the complete chimp genome sequence to decide whether there are significant gene differences caused by this large indel, taking copies into account.

In AC113435 there is a 69-kb gap in the human genome. The chimp sequence opposite the gap is entirely made up of a satellite sequence except for insertion of an L1 fragment. It appears to be a chimp-specific satellite because no search of GenBank or the human genome turns up a similar human sequence. These two long gaps, if real, more than double the length of observed indels in our sample, which was limited to <10 kb. It seems likely that more large indels and regions of rearrangement remain to be observed. Thus the 5% human–chimp difference already published (1) is likely to be an underestimate, possibly by more than a factor of 2. There are a number of rearrangements recognized in chimp vs. human DNA such as the chromosome inversion location recognized in AC006582 (1, 5). One day these will have to be weighed in to calculate the true divergence between human and chimp DNA.

Sea Urchin Polymorphism. Two 52-kb matching regions from the two diploid copies in an individual sea urchin's DNA were sequenced. These sequences will be published along with other data that are part of the project that included the sequencing. Here we just summarize the results as they apply to the issue of the relative number of nucleotides affected by indels compared with base substitutions. The sea urchin has large sequence polymorphism and heterozygosity leading to significant sequence divergence between a diploid pair of DNA sequences. The expected average divergence between the single-copy DNA of two individuals was estimated to be 5% by classical melting curve methods (R.J.B., unpublished data). That compares to the 1.76% observed by similar methods between chimp and human DNA (6, 7). It was also observed that there was a wide range of degrees of divergence among different parts of the sea urchin genome. The comparison of these two sea urchin sequences shows an R_u of 3.0 when they are aligned by the GAPD program described in *Materials and Methods* and $R_u = 3.9$ when they are aligned by a modified Smith Waterman method. By this method there were 1,315 base substitutions and 5,136 nt in indels. These data imply that many more unmatched nucleotides are due to indels than to base substitutions in closely related samples of sea urchin DNA.

Bacterial Strain Sequence Comparison. It appeared of interest to ask whether this observation is restricted to eukaryotes or occurs as well among bacteria. There are several published comparisons of sequences of *E. coli* strains (8–10), but the evolutionary distances are too large to be useful here. For example, K-12 and O15:H7 do not even share a large fraction of protein genes. Comparison of substrains of O15:H7 has been started (9). The regions around *Xba*I sites in the DNA of two substrains of O157:H7 have been examined (11). The conclusion was that more indel events have occurred than base substitutions in this sample of the genomes.

Because of the health crisis caused by *E. coli* strain O157:H7 the genomes of two very closely related substrains have been completely sequenced and are listed by the National Institutes of Health: EDL933 (12) and Sakai (13). As described in *Materials and Methods* the first step in the analysis was to cut the two genome sequences into 100-nt segments and compare each segment with all segments of the DNA from the other substrain. The graphical examination of the phase difference of the best-matching segments showed that in one strain an ≈80-kb section of a genome had been duplicated and inserted in a location 300 kb distant. Furthermore, an ≈400-kb section of one strain had been deleted, inverted, and reinserted in about the location from which it had been deleted (9). This was not the result of a simple event because six regions within it are not present as part of the inversion (i.e., gaps of various sizes). In addition there are at least 34 short regions adding up to 27,500 nt that are inverted in Sakai compared with EDL933, and most occur in very different locations in the two substrains. There are 5,100 nt in 16 short regions of EDL933 that are not recognizable in Sakai. There are many locations where indels are shown by sudden changes in the alignment phase of these segments. The best estimate is that about 4,000 base substitutions have occurred in one or the other substrain. These values suggest a large value of R_u. All of these results are limited by the 100-nt size of the segments used and to some extent confused by repeated sequences. In addition there is a risk that errors in assembly have occurred (9).

The individual long regions lying between the large events of rearrangement were isolated and compared by using GAPD. They each had significant values of R_u, and the length and R_u follow: 1,056,970, 16; 87,493, 195; 719,413, 62; 2,786,791, 73; and 418,655, 1. There is a remarkable amount of variation from region to region, and the weighted average was 55. This must be considered an initial estimate because the nature of the indel events was not examined and the distribution of types of events could be very different from the eukaryotic examples. The very large indels described above as rearrangement events were not included in the calculation of R_u because comparable data are not available for the other species comparisons. It is likely that future comparisons will be made taking into consideration the nature of the individual events. In any case it is clear that indel formation has been the overwhelming process in the divergence of these two substrains of *E. coli* O157:H7.

Discussion

Weber *et al.* (14), reporting a study on human indel polymorphisms, tabulated estimates of the relative number of indels and substitutions in four species. The ratio of indel count to substitution count was 0.59 for *Arabidopsis thaliana* (see below), 0.33 for *Caenorhabditis elegans* (15), 0.19 for *Drosophila melanogaster,* (16) and ranged from 0.22 to 0.43 for human polymorphism measurements (17, 18). Because it is clear that the average indel length is more than a few nucleotides, these data agree with the conclusion of this article. Weber *et al.* (14) did not state, so far as we know, that indels were the major source of unmatched nucleotides between closely related DNAs, although they and other people must have been aware of it.

Cereon corporation (19) has collected many *Arabidopsis* polymorphisms: 37,344 single-nucleotide polymorphisms (SNPs) and 18,579 indels, including 747 large indels deriving from a comparison of two ecotypes of *Arabidopsis thaliana*, Columbia and Landsberg erecta. We had the opportunity of summing the observed indels, multiplying each indel size by the number observed of that length. The total came out to 1,921,866 nt for a sample with 37,344 observed base substitutions. Thus R_u is 51.5. Obviously there is a larger polymorphism difference per nucleotide affected attributable to indels compared with SNPs in this sample because only ≈2% of the unmatched nucleotides in these comparisons are due to base substitutions.

In a study (15) of the comparison of sequences from a Hawaiian isolate to the reference complete sequence of *C. elegans,* 11,000 clones averaging ≈500 nt were sequenced. There were 1,552 indel polymorphisms and 4,670 base substitutions. The longest indel discussed in the paper was only 11 nt, but the authors mention that longer ones were observed. The 2,558 nt in reported indels yield a value of 0.55 for R_u. The *C. elegans* genome contains a number of mobile elements and other repeated sequences and does not seem likely to be immune from the presence of large indels. The work of Robertson (20) comparing the very distant *C. elegans* and *Caenorhabditis brigg-*

Table 2. Summary of indel/substitution ratios

Comparison	R_u
Polymorphism	
A. thaliana	51
C. elegans	(4.2)
D. melanogaster	4.5
S. purpuratus	3–4
E. coli O157:H7	(55?)
Interspecies	
Pan troglodytes/*Homo sapiens*	3.0

R_u is the ratio of unmatched nucleotides attributable to indels to those attributable to base substitutions in the available samples. The *C. elegans* value is put in parentheses because the data are for variation in *mariner* transposon copies. The *E. coli* estimate is uncertain for various reasons mentioned in the text.

sae is not comparable to comparisons of closely related species but does show that large indels are common and an average indel length of 41 nt was observed among recently formed pseudogenes in the *srh* gene family, using conservative criteria for detecting and measuring deletions. Thus it is unlikely that *C. elegans* has fewer long indels than other species even in polymorphic comparisons. Recent work by Witherspoon and Robertson (21) on the indels and base substitutions of *mariner* transposons of *C. elegans* yields an R_u of 4.2, and this is included in Table 2. The average length of indels was 162 nt, and they occurred 0.026 times as often as base substitutions. The authors argue that the neutral evolution of the transposons is a measure of neutral drift in the *C. elegans* genome, and thus this value may be a legitimate estimate for the genome. New data are required for this species before a final R_u can be calculated.

In a study of DNA loss by deletion in *Drosophila* (22) it was found that the number of deletions was ≈0.13 of the number of base substitutions and the average size of the indels was 35 nt (23). Thus R_u is 4.55 for this sample of *D. melanogaster* DNA. A summary of these results is shown in Table 2, and *Drosophila* is similar to the other animals for which there are adequate data.

It is a familiar concept that the bulk of the mutations observed by *Drosophila* genetics is due to mobile element insertions. Presumably mutations including lethals and severely damaging examples will be selected against and many indels may not appear as polymorphic variation or interspecies DNA differences. Thus the observations described here may not be representative of the set of sequence changes as they occur. The same is true for the human gene database, in which lethals are excluded by the methods of collection. To reach the original events as they occur before selection is not easy, but the *Drosophila* mutations offer a possibility.

The evidence supporting the broad conclusions of this paper is bound to be sketchy and incomplete at this early time. However, it is significant. Required for further support are several kinds of evidence: (*i*) confirmation of alignment gaps by PCR or the like; (*ii*) study of more widespread systematic groups; (*iii*) assessment of the size distribution of indels, including long examples; (*iv*) identification of the classes and mechanisms of formation of large numbers of indels; and (*v*) evaluation of the genetic significance of indels.

The conclusion we draw is that indel formation is likely the most rapid and significant form of sequence change (mutation) in eukaryotic evolution and probably bacterial evolution. The mechanism of formation of indels is clear for cases such as insertion of retroposons and other mobile elements, slippage in simple sequence replication, and unequal crossover between similar repeat copies such as *Alu* sequences, leading to deletion of the intervening DNA (24). However in many cases we do not know the mechanism. Human genetic evidence suggests that indels are a major source of gene defects. In one example gene defects affecting the nervous system showed a majority (24/45) that were due to indels (25). Many data suggest that indels are a significant source of evolutionary change.

We thank Dmitri Petrov for help with the *Drosophila* data, Hugh M. Robertson for help with the *C. elegans* measurements, Tanya Berardini for making available *Arabidopsis* data, and Tetsuya Hayashi for help with the *E. coli* comparisons. R.A.C. was supported by Grant IBN-9982875 from the National Science Foundation Developmental Mechanisms Program.

1. Britten, R. J. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 13633–13635.
2. Britten, R. J. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 6148–6150.
3. Ebersberger, I., Metzler, D., Schwarz, C. & Pääbo, S. (2002) *Am. J. Hum. Genet.* **70,** 1490–1497.
4. Gu, X. & Li, W. H. (1995) *J. Mol. Evol.* **40,** 464–473.
5. Nickerson, E. & Nelson, D. L. (1998) *Genomics* **50,** 368–372.
6. Caccone, A. & Powell, J. R. (1989) *Evolution* **43,** 925–942.
7. Springer, M. S., Davidson, E. H. & Britten, R. J. (1992) *J. Mol. Evol.* **34,** 379–382.
8. Welch, R. A., Burland, V., Plunkett, G., III, Redford, P., Roesch, P., Rasko, D., Buckles, E. L., Liou, S. R., Boutin, A., Hackett, J., *et al*. (2002) *Proc. Natl. Acad. Sci. USA.* **99,** 17020–17024.
9. Ohnishi, M., Kurokawa, K. & Hayashi, T. (2001) *Trends Microbiol.* **9,** 481–485.
10. Ohnishi, M., Terajima, J., Kurokawa, K., Nakayama, K., Murata, T., Tamura, K., Ogura, Y., Watanabe, H. & Hayashi, T. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 17043–17048.
11. Kudva, I. T., Evans, P. S., Perna, N. T., Barrett, T. J., Ausubel, F. M., Blattner, F. R. & Calderwood, B. (2002) *J. Bacteriol.* **184,** 1873–1879.
12. Perna, N. T., Plunkett, G., III, Burland, V., Mau, B., Glasner, J. D., Rose, D. J., Mayhew, G. F., Evans, P. S., Gregor, J., Kirkpatrick, H. A., *et al*. (2001) *Nature* **409,** 529–533.
13. Hayashi, T., Makino, K., Ohnishi, M., Kurokawa, K., Ishii, K., Yokoyama, K., Han, C. G., Ohtsubo, E., Nakayama, K., Murata, T., *et al*. (2001) *DNA Res.* **8,** 11–22.
14. Weber, J. L., David, D., Heil, J., Fan, Y., Zhao, C. & Marth, G. 2002) *Am. J. Hum. Genet.* **71,** 854–862.
15. Wicks, S. R., Yeh, R. T., Gish, W. R., Waterson, R. H. & Plasterk, R. H. A. (2001) *Nat. Genet.* **28,** 160–164.
16. Berger, J., Suzuki, T., Senti, K. A., Stubbs, J., Schaffner, G. & Dickson, B. J. (2001) *Nat. Genet.* **29,** 475–481.
17. Antonarakis, S. E., Krawczak, M. & Copper, D. N. (2000) *Eur. J. Pediatr.* **159,** Suppl. 3, S173–S178.
18. Dawson, E., Chen, Y., Hunt, S., Smink, L. J., Hunt, A., Rice, K., Livingston, S., Bupstead, S., Bruskiewich, R., Sham, P., *et al*. (2001) *Genome Res.* **11,** 170–178.
19. Jander, G., Norris, S. R., Rounsley, S. D., Bush, D. F., Levin, I. M. & Last, R. L. (2002) *Plant Physiol.* **129,** 440–450.
20. Robertson, M. H. (2000) *Genome Res*. **10,** 192–203.
21. Witherspoon, D. J. & Robertson, M. H. (2003) *J. Mol. Evol.*, in press.
22. Petrov, D. A. (2002) *Genetica* **115,** 81–91.
23. Petrov, D. A. (2002) *Theor. Popul. Biol.* **61,** 533–546.
24. Lehrman, M. A., Russell, D. W., Goldstein, J. L. & Brown, M. S. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 3679–3683.
25. Mockus, S. M. & Vrana, K. E. (2000) in *Genetic Polymorphisms and Susceptibility to Disease*, eds. Miller, M. S. & Cronin, M. T. (Taylor and Francis, New York), p. 231.