Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels

Roy J. Britten*

California Institute of Technology, 101 Dahlia Avenue, Corona del Mar, CA 92625

Contributed by Roy J. Britten, August 22, 2002

Five chimpanzee bacterial artificial chromosome (BAC) sequences (described in GenBank) have been compared with the best matching regions of the human genome sequence to assay the amount and kind of DNA divergence. The conclusion is the old saw that we share 98.5% of our DNA sequence with chimpanzee is probably in error. For this sample, a better estimate would be that 95% of the base pairs are exactly shared between chimpanzee and human DNA. In this sample of 779 kb, the divergence due to base substitution is 1.4%, and there is an additional 3.4% difference due to the presence of indels. The gaps in alignment are present in about equal amounts in the chimp and human sequences. They occur equally in repeated and nonrepeated sequences, as detected by REPEATMASKER (http://ftp.genome.washington.edu/RM/RepeatMasker.html).

base substitutions | insertion/deletion differences | DNA evolution

Many years ago, the hydroxyapatite method for measuring sequence divergence between species was developed by Dave Kohne and me. In this method, hybrid DNA strand pairs were formed from small fragments and the temperature at which they were disassociated determined. This method was used by several groups to compare chimpanzee and human DNA (1–3), and the best measurements suggested a divergence of 1.76% of single-copy DNA preparations. This observation led to the widespread quotations that we were 98.5% similar to chimps in our DNA, sometimes, mistakenly, that we had 98.5% gene similarity. I have compared the actual DNA sequences for the chimp BACs that have become available. Previously, two groups have measured the divergence between human and chimp DNA sequences to be 1.25% for base substitutions in about 2 Mb of primarily single-copy DNA (4, 5). Another estimate of the divergence was 1.23% for sequenced ends of many chimpanzee BACs (6). These groups did not compute the insertion/deletion contribution to divergence. For the work reported here, there are about 779 kb of BAC sequences in GenBank that can now be aligned with the human genome sequence, although that number will increase, because the full chimpanzee sequence will become available in due course.

Methods

The chimpanzee BAC sequences (K. C. Worley and colleagues and L. Yang and colleagues; see accession nos. listed in Table 1) were downloaded from GenBank, and the "BLAST the human genome" facility was used to identify the regions of alignment. The available sequence alignment programs are not good at comparing long regions and use heuristic methods to establish gaps in alignment, leading to some uncertainty. Therefore, a FORTRAN program was written to ascertain the actual gaps in alignment. For this purpose, the matching human sequence is trimmed at the beginning to start exactly in alignment with the chimp sequence. The program moves sequentially along the pair of sequences. When it detects a mismatch, if the next 20-nt match is in more than 15 positions, it is counted as a substitution. If the next 20 nt do not match this well, then the program looks for a nearby region that does match and tests for gaps up to 5 kb long in either sequence to see whether a 20-nt-long test region matches more than 15 nt. It is a fairly specific program, working well for very similar sequences, which do not have gaps larger than 5 kb; this is true for the examples in Table 1. It can be adjusted for sequence pairs with more divergence and larger gaps and works for comparison with human sequences of examples of baboon BACs that have been sequenced.

Results

Table 1 summarizes the human/chimp mismatch seen in the five BACs examined over the lengths that could be easily aligned. The first example is listed in two rows to represent the two different regions of chromosome 12 that accurately aligned with parts of the chimp sequence. Table 1, column 4, shows the percent base substitution, which, on average, is only slightly higher that the previous estimates for the whole single-copy fraction of the genome. There is considerable variation, ranging from 1.2 to 1.69%, indicating what are apparently regional differences in the degree of divergence. The compared parts of the BACs average 127 kb each, and the number of substitutions was about 1,500; thus, the expected small number statistical fluctuation does not account for the observed variation.

Table 1, column 5, shows the percent of these alignment lengths that were in gaps in either the human or the chimp sequences. On average, it is much more than twice as great as the base substitution percent. It is these values that have made the total divergence large compared with the old estimates. It appears appropriate to me to consider the full length of the gaps in estimating the interspecies divergence. These stretches of DNA are actually absent from one and present in the other genome. In the past, indels have often simply been counted regardless of length and added to the base substitution count, because that is convenient for phylogenetics.

The Size of the Indels. The length distribution of the gaps is given in Tables 2 and 3. Table 2 shows the smaller gaps for which given sizes occur multiple times in this sample. The second column shows the number of occurrences. Single nucleotide gaps occur with the largest frequency, and the frequency falls monotonically for larger gaps, except for small number fluctuation. Table 3 shows the larger gaps, each of which occurs once in the sample. The fall in frequency with length appears to continue among these cases as their spacing steadily increases with length. The frequency falls somewhat faster than in proportion to the length of the indels but not as fast as the square of the length.

Mechanism of Formation of the Indels. Only a few interesting cases have been examined, and it appears that several mechanisms contribute to the observed indels. A number of gaps occur in homopolymers (stretches of identical nucleotides). In fact, there are only 10 such homopolymer regions longer than 10 nt in one of the BACs that was examined, and all but two of these homopolymers have gaps. The chance of slippage seems to be large for these homopolymer regions, with an approximate 80% occurrence of gaps in the short evolutionary distance between

Abbreviations: BAC, bacterial artificial chromosome; NPH, normalized percent hybridization.

*E-mail: rbritten@cco.caltech.edu.

Table 1. Divergence of the sample regions

Name	Length,* nt	Chrom§	Substitution,† %	Indels,‡ %	Total, %
AC006582	72,176	12q14	1.20	0.63	1.87
AC006582	93,769	12p12	1.40	4.12	5.52
AC007214	132,974	12q13	1.31	3.90	5.21
AC097335	148,994	2q14	1.69	3.58	5.27
AC096630	150,877	20p11.2	1.58	2.51	4.09
AC093572	180,352	22q11	1.22	4.50	5.72
Average¶			1.41	3.91	4.85

The total length of these aligned regions is 779,142 bp.
*The length of the region compared.
†The percent of nucleotides replaced by a different nucleotide.
‡The sum of the length of all gaps in both human and chimpanzee sequences as percent of aligned chimp length.
§Chromosomal location listed by ''Search the Human Genome.''
¶Weighted (by length) average.

chimp and human. This slippage is apparently typical, because observations of mononucleotide microsattelites in a 5.1-mb comparison between chimp and human (7) showed just 25% were identical in length. The largest gap observed is 4,263 nt long in AC007214, missing from the human sequence. In chimp, the region is primarily made up of L1 repeat segments that occur at each end. Together, the L1 repeats constitute 60% of the length of the region, which includes, in addition, a MaLR element 382 nt long. It appears that, in this case, an unequal crossover between parts of two L1 elements has led to a large deletion, including a non-L1 region.

AC007214 also contains a 280-nt-long element missing from the chimp sequence, also identified by REPEATMASKER (http://ftp.genome.washington.edu/RM/RepeatMasker.html) as part of an L1 element. There are two distant full length divergent copies of this sequence in this aligned region. It appears that an insertion or deletion event of an L1 repeated sequence has occurred. AC0093572 contains six indels greater than 150 nt, and they have been examined with REPEATMASKER. All of them contain repeated elements, although in many cases they do not extend to the termini. One indel contains six Alu repeats, and two contain SVA repeats and no other sequences. AC096630 has four indels greater than 150 nt, and one of these consists of five Alu repeats and little other sequence except for a terminal 42-nt-long poly (GT). Three of the indels contain no identifiable repeated sequences. AC097335 has only two indels greater than 150 nt, and one is primarily an Alu repeat. The other is not a recognized repeat except for an internal 30-nt-long AT-rich low-complexity sequence. Thus, no consistent single cause can be identified for even the longer indels, and several mechanisms must be responsible for the observed insertions and deletions. Presumably, a number of the gaps are due to slippage in micro- and minisatellites, although no systematic search was done. In a comparison of 5.1 Mb of chimp and human DNA for microsattelites, 1,174 were observed to be different in length (7), suggesting that 169 of the gaps on Tables 2 and 3 could have been due to microsattelite length differences.

Table 2. Lengths of indels

Length, nt	No.	Length, nt	No.
1	410	17	6
2	148	18	4
3	108	19	1
4	96	20	3
5	33	21	3
6	33	22	1
7	19	23	1
8	28	24	5
9	12	25	2
10	18	26	2
11	9	27	1
12	14	28	1
13	7	30	1
14	8	31	3
15	3		
16	4		

Table 3. The longer indels that occur once in the sample

Length, nt
33, 34, 35, 36, 37, 39, 44, 45, 52, 57, 60, 64, 68, 70, 76, 78, 84, 122, 134, 161, 260, 280, 317, 503, 914, 920, 927, 929, 932, 1,459, 1,480, 2,147, 2,933, 3,545, 4,263

Background

Percent Nucleotide Substitution. Estimates were made of the sequence divergence of the total ''single-copy'' DNA prepared by fractionation after partial reassociation to remove repeated sequences (1–3). Caccone and Powell (8) discuss the normalized percent hybridization (NPH) and favor using T_m reduction without correction for extent of hybridization. Their value is 1.59° between chimpanzee and human. Using a modern value (9) of 1.11% sequence divergence per degree T_m reduction, the divergence converts to 1.76%, which is in poor agreement with the average value of 1.4% in Table 1. The other measurements (1–3) of T_m reduction of chimp/human hybrid strand pairs are in reasonable agreement with the Caccone–Powell measurement. That is, all are larger than the value on Table 1. A large part of the difference is presumably due to the effect of indels on the melting temperature of sequence hybrids.

Modern measurements by direct sequencing indicate a divergence due solely to base substitution in single-copy DNA of 1.25% (4, 5). Part of the reason our estimate of 1.4% is larger is because repeated sequences are included, and they diverge somewhat more than typical single-copy sequences (4). Some of this difference is due to the CpGs present in Alu repeats, which are mutated at 10 times the rate of other nucleotides. One of the BACs was checked for CpGs and, as expected, a larger fraction of the CpGs showed base substitutions. In addition, the ends of many chimpanzee BACs have been sequenced and compared with the human genome (6). The average divergence was 1.23% for the 19,568,934 good sequences that could be matched, agreeing well with the other single-copy sequence comparisons.

Fraction of Indels. Caccone and Powell (8) present a curve of NPH vs. T_m reduction that suggests the failure to hybridize for chimp/human compared with human/human is about 5%. However, the tabulated data are not convincing, with a range of 61–113% NPH and an average of 89%. The data of Sibley *et al.* are comparable, and they did not list the NPH. In ref. 3, they do print actual melting curves, and the NPH can be picked off the graphs: it ranges from 92 to 99% and averages about 95%. All measurements agree there is a failure to hybridize that increases with T_m reduction or distance between the species being compared. The origin of the failure to hybridize has never been clear. Now it seems that much of it is due to actual missing DNA sequences resulting from indel events. Some part of it may also have been due to local regions of the genomes showing larger degrees of base substitution. Complete genome comparisons will settle these issues. These comparisons should be particularly interesting for *Drosophila* species, which show large reduction in NPH compared with T_m reduction (10).

Rearrangement Events. The similar regions in the human genome to BAC AC006582 are in two distant parts of chromosome 12. There is a clean break between them without overlap or gap. If we accept these sequences as being correct, it appears that there was a rearrangement event in either the chimp or human genome. One BAC clone that was only terminally sequenced (6) found matching regions in human 12q15 for one end and 12p12 for the other end. The authors identify this clone as crossing a chromosome inversion location known between chimpanzee chromosome 10 and human chromosome 12 (11). In the case of AC007214, there is an internal 23-kb sequence that is in large part 99% similar to an internal region of AC006582. It appears there has been a duplication of a stretch of the chimp genomic DNA sequence, and thus the aligned region of AC007214 was terminated to avoid a duplication of the regions counted in the divergence data. The suggestion is that rearrangement events have been frequent in the evolution of these primate DNA sequences. It seems quite unlikely, but not impossible, that imperfections in the draft genome sequences are responsible for the appearance of rearrangement events.

Discussion

This is an observation of the major way in which the genomes of closely related primates diverge—by insertion/deletion. More nucleotides are included in insertion/deletion events (3.4%) than base substitutions (1.4%) by much more than a factor of two. However, the number of events is small in comparison. About 1,000 indels listed in Tables 2 and 3 compared with about 10,000 base substitution events in this comparison of 779,142 nt between chimp and human. Little can be said about the effect of these indel events. There were so few gene regions in this small sample that a statistical analysis of their occurrence did not seem worthwhile. That will have to wait until larger regions for comparison become available.

The gaps should be useful markers for distinguishing the evolutionary relationship among closely related species, because once a deletion occurs, it would be very unlikely that the missing region sequence would be reconstructed *de novo*, unless it was repeat that could be copied. If a set of these indels were collected, they could be tested for their presence in gorilla by PCR and ultimately resolve the human–chimp–gorilla trichotomy in an unarguable way. However, it now seems almost certain that chimp or bonobo is our nearest relative (12, 13).

One interesting observation is that the sequence divergence between chimp and human is quite large, in excess of 20% for a few regions. Some of the larger gaps are broken by regions within them that align with appropriate segments of the other species' DNA sequence but only have distant similarity. These observations suggest that complex processes, presumably involving repeated sequences and possible conversion events, may occur that will require detailed study to understand. The uncertainty in the estimate of 3.4% indels on Table 1 cannot be directly evaluated. In the first place, the sample of 779 kb is small, and the variation between the different BACs is large. Further, there may be large gaps that were missed as part of chimpanzee BAC sequences that could not be aligned with the human genome. Nevertheless, the conclusion is clear that comparison of the DNA sequences of closely related species reflects many events of insertion and deletion. It is the result of a major evolutionary process.

I thank John Williams for help in data processing and DNA sequence comparisons.

1. Hoyer, B. H., De Velde, N. W., Goodman, M. & Roberts, R. B. (1972) *J. Hum. Evol.* **1,** 645–649.
2. Sibley, C. G. & Ahlquist, J. E. (1984) *J. Mol. Evol.* **20,** 2–15.
3. Sibley, C. G., Comstock, J. A. & Ahlquist, J. E. (1990) *J. Mol. Evol.* **30,** 202–236.
4. Ebersberger, I., Metzler, D., Schwarz, C. & Paabo, S. (2002) *Am. J. Hum. Genet.* **70,** 1490–1497.
5. Chen, F.-C. & Li, W.-H. (2001) *Am. J. Hum. Genet.* **68,** 444–456.
6. Fujiyama, A., Watanabe, H., Toyoda, A., Taylor, T. D., Itoh, T., Tsai, S.-F., Park, H.-S., Yaspo, M.-L., Lehrach, H., Chen, Z., *et al.* (2002) *Science* **295,** 131–134.
7. Webster, M. T., Smith, N. G. C. & Ellegren H. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 8748–8753.
8. Caccone, A. & Powell, J. R. (1989) *Evolution (Lawrence, KS)* **43,** 925–942.
9. Springer, M. S., Davidson, E. H. & Britten, R. J. (1992) *J. Mol. Evol.* **34,** 379–382.
10. Caccone, A. & Powell, J. R. (1990) *J. Mol. Evol.* **30,** 273–280.
11. Nickerson, E. & Nelson, D. L. (1998) *Genomics* **50,** 368–372.
12. Wimmer, R., Kirsch, S., Rappold, G. A. & Schempp, W. (2002) *Chromosome Res.* **10,** 55–61.
13. Ruvolo, M. (1997) *Mol. Biol. Evol.* **14,** 248–265.