

# The History of Nucleotide Substitution Pattern in Human

P. F. Arndt

Max-Planck Institute for Molecular Genetics, Ihnestr. 73, 14195 Berlin, Germany, E-mail: arndt@molgen.mpg.de

Received November 3, 2003; in final form, December 22, 2003

**Abstract**—Substantial regional variations of substitutional processes have recently been reported from human/mouse comparisons. However, several features including the C+G dependence and the CpG-based transition effect remain obscure. Utilizing the vast amount of transposable elements in the human genome, we performed detailed analysis of the substitutional and insertion/deletion patterns along the human lineage in a regional and time-resolved fashion. We observed a drastic increase in the CpG-based transition frequency at about the time of the mammalian radiation. We also observed clear regional biases of substitution patterns, most notably a bias to C+G enrichment towards the telomeres.

**Key words:** nucleotide substitution, CpG methylation

## INTRODUCTION

A comparative study of the human and mouse genomes can in principle provide a lot of information about the evolutionary history of these genomes. Recently, Hardison *et al.* has presented a study of this kind, which started from a whole genome alignment of the human genome [2] and the mouse genome [3] build by the BLASTZ alignment program [4]. To estimate the evolutionary change between human and mouse, homologous ancestral repetitive elements (REs) have been aligned and the observed base substitutional activity has been mapped along the chromosomes in windows of 5 Mb. This way, substantial variations of six different measures of evolutionary change have been found.

Here we want to present work that will complement the understanding of regional variations in the rates of neutrally evolving DNA sequence in human. In contrast to the above-mentioned comparative study, our analysis is based solely on the human genome. We utilize the vast amount of repetitive sequence (about 50% of the human genome) as “fossil record” to extract information about the substitutional process acting on neutrally evolving DNA. We primarily include data of young REs into our analysis. Because there are a lot more (~5×) human-specific REs than the ancestral REs that survived in both human and mouse lineages, and because the younger REs can be much

better aligned, we are able to collect more detailed information. Further, our analysis does not suffer from potential distortion of the estimated substitution frequencies caused by an elevated rate of substitutions solely in the mouse lineage. We are able to measure *regional substitutional patterns* (on a 1-Mb scale along the chromosomes), and by combining information from differently old families of REs we are able to reconstruct the *evolutionary history* of the human genome for the last ~250 Myr.

The precision of our method is greatly enhanced by the explicit incorporation of the neighbor-dependent CpG-methylation–deamination process, which is known to be the predominant substitution process in vertebrates [5–7]. In contrast to neighbor-independent nucleotide substitutions, e.g., transversions and transitions, the dynamics of *neighbor-dependent* substitutions such as the methyl-CpG assisted transition (CpG → CpA/TpG) is much more complex to analyze. It is commonly (and erroneously) assumed that the CpG-based transition only affects CpG dinucleotides, and one can learn about all the other substitution processes by excluding CpG sites from the analysis. This is based implicitly on the view that the CpG process is a small perturbation to the neighbor-independent substitution model. However, the rate of the CpG-based transition is actually estimated to be as high as 40 times that of a transversion [7]. Hence

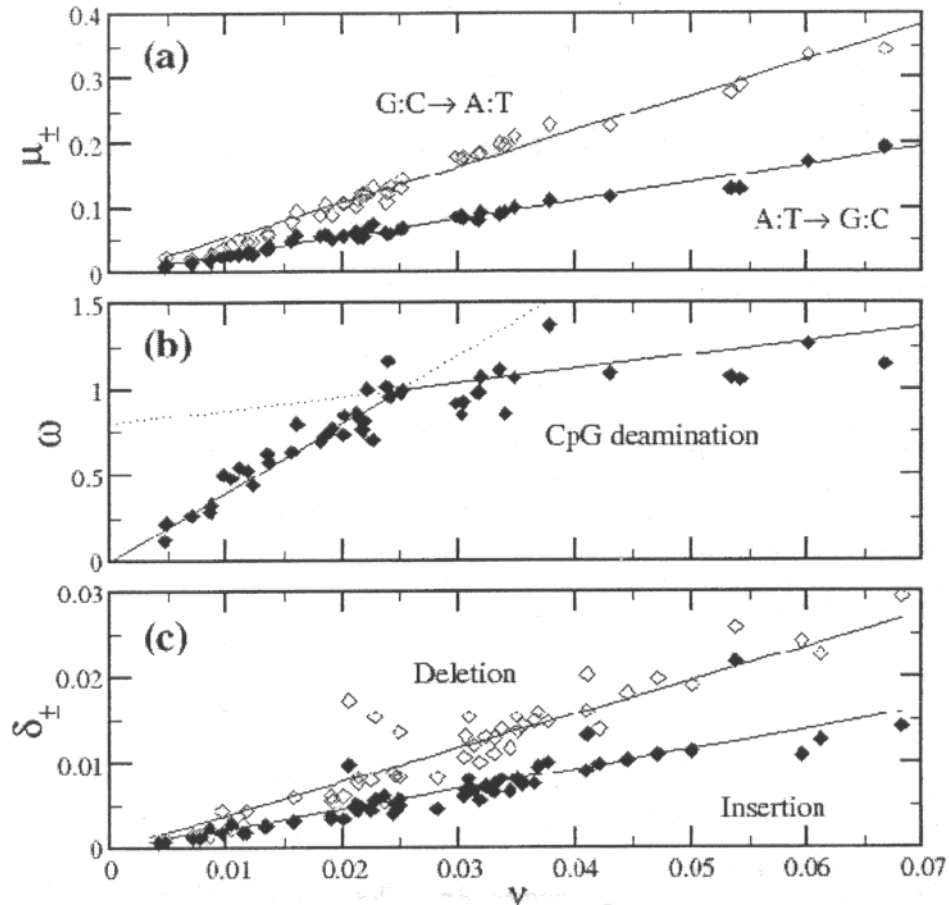


Fig. 1. (a) Transition frequencies, (b) CpG-deamination frequency, and (c) insertion and deletion frequencies estimated from several repetitive elements as a function of their age  $v$ .

CpG's in the ancestral sequence very rapidly decay into CpA or TpG, which are subsequently mutated into other bases. Consequently, the existence of the CpG process affects many sites that do not directly involve CpG's.

For our study we extracted from the human genome numerous copies of the commonly encountered families of repeats, the SINEs (Alu, MIR) and LINES (L1, L2, L3) [9-12]; details of this procedure are described by Arndt *et al.* [7]. Thus our analysis is based on as much as 15% of the available human genome data. In the sequel we analyzed the observed substitutions in the individual copies of REs to estimate the four transversion, two transition frequencies, and the frequency of the CpG-assisted transition, where frequencies denote numbers of substitutions per base pair after insertion of the RE into the genome. For this purpose we used a maximum likelihood (ML) approach, which also takes into account effects due to multiple and back-substitutions.

## RESULTS AND DISCUSSION

### Genome-Wide Mutation Pattern

**Transitions and transversions.** In Fig. 1a, we plot the two neighbor-independent transition frequencies, (denoted by  $\mu_{\pm}$ ) against the *average* of the four (rather similar) transversion frequencies (denoted by  $v$ ) for each subfamily of REs. From this figure, we observe first that the two neighbor-independent transition frequencies show a remarkably linear dependence ( $R = 0.99$ ) on the average transversion frequencies. This suggests that the genome-wide averaged neighbor-independent substitution pattern has not changed since the time the oldest elements (L3, L2, MIR) entered the genome. Fitting these two transition frequencies to straight lines, and identifying the slopes ( $\mu_{\pm}/v$ ) as the relative transition rates, we find  $\mu_{+} = (2.74 \pm 0.04)v$ , and  $\mu_{-} = (5.5 \pm 0.1)v$ . [An analysis based on a few families of DNA transposons has already been performed by Lander *et al.* [2], yielding estimates  $\mu_{+} \approx 2.5 v$ , and  $\mu_{-} \approx 5 v$ .]

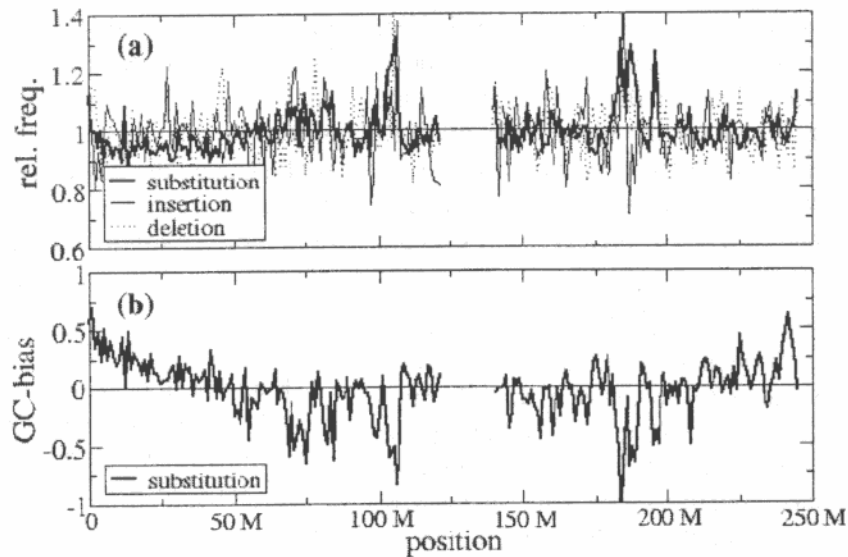


Fig. 2. (a) Regional substitution, insertion, and deletion frequencies; and (b) GC-bias of substitutions along the human chromosome 1.

Given the linearity of the data in Fig. 1a, it is convenient to use the horizontal axis as the “time” axis. Calibrating the time scale using estimates of the absolute insertion time of the different Alu subfamilies [13], we find each unit of  $\nu = 0.01$  in the average transversion frequency to correspond to approximately 35 Myr, with the entire dataset spanning nearly 250 Myr. Thus our analysis reveals that the same substitution pattern has been maintained for the past 250 Myr, much before the period of mammalian radiation that occurred 80–100 Myr ago.

**CpG-based transition.** The corrected CpG-based transition frequencies,  $\omega$ , shown in Fig. 1b present a big surprise: the data clearly present two regimes characterized by very different slopes. This finding is supported by analyzing REs across different families of SINEs and LINEs, and we verified that this finding is not an artifact of the ML analysis [7]. To quantitate the extent to which the transition rates had changed in the past, we divided the data into the two sets of “young” and “old” REs with respect to a threshold value  $\nu_0$ , with the threshold adjusted so that the sum of the squared residuals of linear regressions to the data in both sets would be minimal. This minimum was found for  $\nu_0 = 0.025$ , with the slope of  $39.5 \pm 2.6$  for the young elements and  $8.4 \pm 2.5$  for the old elements. It is natural to identify the two slopes with the relative rates  $\omega/\nu$  before and after  $\nu_0$  (as justified by Arndt *et al.* [7]). This leads to the conclusion that a 4- to 8-fold increase in the CpG-based

transition rate occurred at  $\nu_0 \approx 0.025$  or  $\sim 90$  Myr ago, corresponding roughly to the time of the mammalian radiation [14–16]. This conclusion is corroborated by another independent observation by Arndt *et al.* [7].

**Insertion and deletions.** We repeated the analysis to study insertions and deletions using the different RE families. For each gapped alignment of a RE with its master sequence, we collected separately the number of insertion and deletion events, as well as the length of each inserted or deleted segment. The insertion and deletion frequency per nucleotide (denoted by  $\delta_+$  and  $\delta_-$  respectively) were computed for each RE subfamily and plotted against the average transversion frequencies in Fig. 1c. One can see that both the insertion and deletion rates have remained remarkably constant over the past 250 Myr, with  $\delta_+/\nu \approx 0.23$  and  $\delta_-/\nu \approx 0.40$ . This result is consistent with the qualitative finding that the deletion rate is approximately twice the insertion rate reported by Waterston *et al.* [3]. The variation in the amount of deletions seems to be larger than for the insertions. It remains to be investigated whether this variation is due to RE specific effects and is a signature of a process that deletes specific small motifs from the human genome.

### Regional Mutation Patterns

The abundance of repetitive elements allows us to also estimate the regional substitution patterns along each chromosome. As mentioned above, regional variations in the frequencies of substitutions

have already been found by Hardison *et al.* [1]. Here we can perform a similar regional analysis for mutations along the human lineage alone, by repeating the analysis shown in Fig. 1 for the human-specific REs residing in each genomic region. In this way, we are able to collect detailed information (i.e., all 7 substitution frequencies and indel frequencies) at a resolution of 1 Mb.

In Fig. 2a, we show regional substitution, insertions, and deletion frequencies. The regional substitution frequency is taken to be the sum of all transversion and transition frequencies (excluding the CpG-based process) for each 1-Mb window regional substitution frequency relative to the genomic average along chromosome 1. For this analysis we took only the Alu elements into account, since they are the most recent insertions and therefore reflect best the current substitutional activity along the human chromosomes. Further, taking the age of each Alu sub-family into account, we build an averaged substitution pattern in each region. (The observations reported here are typical of those exhibited by the other autosomes.) While the variation is confined to the range of  $\pm 10\%$  in most regions, we see distinct substitutional "hot spots" localized to regions of only a few Mb in length with more than 30% excess activity. The variation in substitution frequency is echoed by variations in the insertion and deletion frequencies, especially in the vicinity of the hot spots.

In addition, we study the influence of the substitutional process on large-scale base compositional variations of the human genome, which are known as genomic isochores [17, 18]. The origin, timing, and implications of the human isochore structure are still controversial (see Eyre-Walker and Hurst [19] for a review). Here, we examined regional substitutional biases, i.e., tendency for substitutions to be GC-enriching or GC-depleting. We plotted in Fig. 2b the difference between the sum of the GC-enriching and GC-depleting substitutions, relative to and normalized by its genomic averaged value, for each 1-Mb window. Distinct GC-enriching biases can be seen for the two telomeric regions, with the individual substitution frequencies changing by as much as 20–30%.

This analysis can be repeated for each available mammalian and vertebrate genome. Together they will reveal a very accurate regional and time-resolved picture of the evolution of mammalian genomes. In

the future, detailed information about the background process responsible for nucleotide substitution as well as base insertions and deletions will also help address an outstanding problem facing comparative genomic analysis, i.e., the difficulty of distinguishing sequence homology due to functional constraints from that due to common evolutionary ancestry.

## REFERENCES

1. Hardison, R.C., *et al.*, Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution, *Genome Res.*, 2003, vol. 13, no. 1, pp. 13–26.
2. Lander, E.S., *et al.*, Initial sequencing and analysis of the human genome, *Nature*, 2001, vol. 409, no. 6822, pp. 860–921.
3. Waterston, R.H., *et al.*, Initial sequencing and comparative analysis of the mouse genome, *Nature*, 2002, vol. 420, no. 6915, pp. 520–562.
4. Schwartz, S., *et al.*, Human-mouse alignments with BLASTZ, *Genome Res.*, 2003, vol. 13, no. 1, pp. 103–107.
5. Hess, S.T., Blake, J.D., and Blake, R.D., Wide variations in neighbor-dependent substitution rates, *J. Mol. Biol.*, 1994, vol. 236, no. 4, pp. 1022–1033.
6. Arndt, P.F., Burge, C.B., and Hwa, T., DNA Sequence Evolution with Neighbor-Dependent Mutation. in 6th Annual International Conference on Computational Biology RECOMB2002, 2002, Washington DC: ACM Press.
7. Arndt, P.F., Petrov, D.A., and Hwa, T., Distinct changes of genomic biases in nucleotide substitution at the time of Mammalian radiation, *Mol. Biol. Evol.*, 2003, vol. 20, no. 11, pp. 1887–1896.
8. Lio, P. and Goldman, N., Models of molecular evolution and phylogeny, *Genome Res.*, 1998, vol. 8, no. 12, p. 1233–1244.
9. Jurka, J. and Milosavljevic, A., Reconstruction and analysis of human Alu genes, *J. Mol. Evol.*, 1991, vol. 32, no. 2, pp. 105–1021.
10. Smit, A.F. and Riggs, A.D., MIRs are classic, tRNA-derived SINES that amplified before the mammalian radiation, *Nucleic Acids Res.*, 1995, vol. 23, no. 1, pp. 98–102.
11. Smit, A.F., *et al.*, Ancestral, mammalian-wide sub-families of LINE-1 repetitive sequences, *J. Mol. Biol.*, 1995, vol. 246, no. 3, pp. 401–417.

12. Jurka, J., Repbase update: a database and an electronic journal of repetitive elements, *Trends Genet.*, 2000, vol. 16, no. 9, pp. 418–420.
13. Kapitonov, V. and Jurka, J., The age of Alu subfamilies, *J. Mol. Evol.*, 1996, vol. 42, no. 1, pp. 59–65.
14. Kumar, S. and Hedges, S.B., A molecular timescale for vertebrate evolution, *Nature*, 1998, vol. 392, no. 6679, pp. 917–920.
15. Eastal, S., Molecular evidence for the early divergence of placental mammals, *Bioessays*, 1999, vol. 21, no. 12, pp. 1052–1058; discussion 1059.
16. Murphy, W.J., *et al.*, Molecular phylogenetics and the origins of placental mammals, *Nature*, 2001, vol. 409, no. 6820, pp. 614–618.
17. Filipski, J., Thiery, J.P., and Bernardi, G., An analysis of the bovine genome by Cs<sub>2</sub>SO<sub>4</sub>-Ag density gradient centrifugation, *J. Mol. Biol.*, 1973, vol. 80, no. 1, pp. 177–197.
18. Bernardi, G., Isochores and the evolutionary genomics of vertebrates, *Gene*, 2000, vol. 241, no. 1, pp. 3–17.
19. Eyre-Walker, A. and Hurst, L.D., The evolution of isochores, *Nat. Rev. Genet.*, 2001, vol. 2, no. 7, pp. 549–555.