BIOINFORMATICS

Vol. 20 no. 10 2004, pages 1482–1485
doi:10.1093/bioinformatics/bth105

Regional and time-resolved mutation patterns of the human genome

Peter F. Arndt[1,*] and Terence Hwa[2]

[1]Max Planck Institute for Molecular Genetics, 14195 Berlin, Germany and [2]Physics Department, University of California at San Diego, La Jolla, CA 92093, USA

Received on October 12, 2003; accepted on February 3, 2004

ABSTRACT

Motivation: 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.

Results: 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 enrich the C+G content toward the telomeres.

Availability: The programs used are available upon request from the authors.

Contact: peter.arndt@molgen.mpg.de

1 INTRODUCTION

A comparative study of the human and mouse genome can in principle provide a lot of information about the evolutionary history of these genomes. Recently, Hardison *et al*. (2003) presented a study of this kind, that started from a whole genome alignment of the human genome (Lander *et al*., 2001) and the mouse genome (Waterston *et al*., 2002) build by the BLASTZ alignment program (Schwartz *et al*., 2003). To estimate evolutionary change between human and mouse, homologous ancestral repetitive elements (REs) have been aligned and the observed base substitutional activity have been mapped along the chromosomes in windows of 5 Mb. This way, substantial variation of six different measures of evolutionary change has been found.

Here, we present complementary work that is based solely on the human genome. We utilize the vast amount of repetitive sequence (about 50% of the human genome) as a 'fossil record' to extract information about the local substitutional process. Most of the REs reside in the intergenic regions and are in general functionally neutral, so that the substitution patterns estimated are those of non-functional sequence. We primarily include data of young REs into our analysis. There are a lot more ($\sim 5\times$) human-specific REs than the ancestral REs (i.e. those that survived in both the human and mouse lineages) used in the comparative method, and the younger REs can be better aligned to their respective master sequences. Furthermore, analysis based on human-specific REs avoids potential distortions due to variations of substitutions solely in the mouse linage. We are able to measure regional substitutional frequencies and biases on a 1 Mb scale along the chromosomes. By combining information from various old families of REs, we are able to examine quantitatively the substitutional history of the human genome for the last $\sim$250 million years (Myr). The substitution pattern we obtained shows interesting regional and temporal variations and may shed light on the origins and timing of the large-scale variations in base composition along the chromosomes known as genomic isochores (Bernardi, 2000).

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 (Hess *et al*., 1994; Arndt *et al*., 2002, 2003). In contrast to neighbor-independent nucleotide substitutions, e.g. transitions and transversions (Lio and Goldman, 1998), the dynamics of neighbor-dependent substitutions such as the methyl-CpG-assisted transition (CpG $\rightarrow$ 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 (neighbor-independent) substitution processes simply by excluding CpG sites from the analysis. This assumption is based implicitly on the view that the CpG process is a small perturbation to the neighbor-independent substitutions. However, the rate of the CpG-based transition is actually estimated to be as high as $40\times$ that of a transversion (Arndt *et al*., 2003). With such high a rate, CpGs in the ancestral sequence decay very rapidly into CpA or TpG, which are subsequently mutated into other bases. Consequently, the existence of the CpG process affects many sites that do not apparently involve CpGs in the observed sequences.

*To whom correspondence should be addressed.

Bioinformatics 20(10) © Oxford University Press 2004; all rights reserved.

We present an extension of our previous analysis (Arndt *et al*., 2003) and study regional variations of nucleotide substitution pattern along the chromosomes in windows of 1 Mb. For completeness, we also present results on the genome-wide substitution pattern and include results on nucleotide insertions and deletions.

2 MATERIALS AND METHODS

Using RepeatMasker (http://repeatmasker.genome.washington.edu), we extracted from the human genome (Build 28) numerous copies of the commonly encountered repetitive elements (Jurka, 2000). With this procedure about 46% of the human genome was identified to be either part of a RE or a sequence of low complexity. The subsequent analysis was focused on elements from 42 subfamilies, the SINEs (Alu, MIR) and LINEs (L1, L2, L3) (Jurka and Milosavljevic, 1991; Smit and Riggs, 1995; Smit *et al*., 1995), which comprise about 410 Mb of sequence data (∼12% of the human genome).

In the analysis of our data the 'star' phylogeny was assumed for each subfamily of REs. This assumption is based on the known biology of retrotransposons and supported by previous phylogenetic analysis of the transposons (Britten *et al*., 1988; Jurka and Smith, 1988; Jurka and Milosavljevic, 1991). The assumption would be invalidated if a significant fraction of the REs is generated by duplication. However, the latter is estimated to affect under 10% of the REs (J.Jurka, personal communication), making the star phylogeny a reasonable starting point. Given the low degree of sequence divergence of the copies of REs, we assume that each copy evolves as a unique sequence. Further, the extremely low divergence of the youngest REs (Jurka, 2000) in the human genome suggests that retrotranscription errors can be ignored safely.

To estimate substitutional frequencies for each subfamily of REs, we used a maximum-likelihood (ML) approach. The observed data are given by the pair-wise gapped alignments of each identified copy of a RE and its ancestral sequence from the RepBase (Jurka, 2000). To implement the ML approach, we need to specify a substitution model. We chose a very general model comprising all possible single nucleotide transversions (8) and transitions (4) as well as the CpG-based transition CpG → CpA/TpG. By design, we also include multiple- and back-substitutions into our model. Each process and its complementary process are assumed to occur with the same substitution frequency per site. Hence the model is parametrized by seven substitution frequencies. The likelihood of observing the given alignment data given a particular model can easily be computed (Arndt *et al*., 2002, 2003) and maximized adjusting the substitution frequencies using standard algorithms (Press *et al*., 1992). The typical error of the ML method is estimated by bootstrap (Press *et al*., 1992). Due to the large amount of sequence data (especially for the genome-wide analysis), the estimated errors are small and will be omitted in the following.

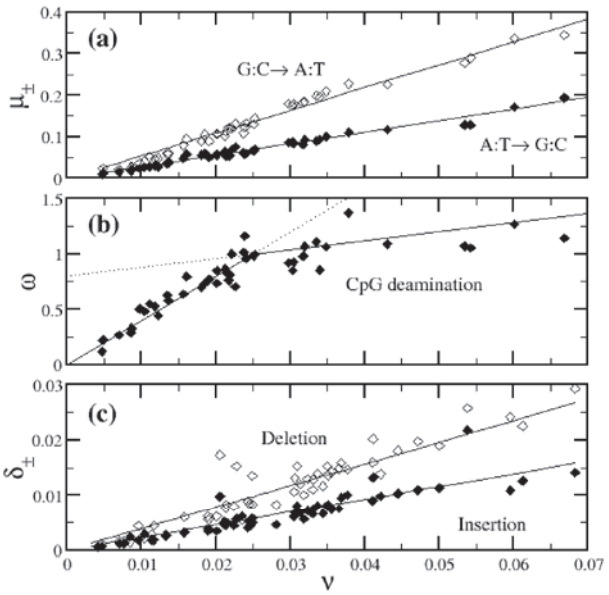

Fig. 1. (**a**) Transition frequencies, (**b**) CpG-deamination frequency and (**c**) insertion and deletion frequencies estimated from several repetitive elements as a function of their 'age', ν.

The performance of this method had been tested extensively using synthetically aged sequences with known substitution frequencies starting from a known ancestral sequence (Arndt *et al*., 2003). We found that given the amount of sequence data we have, substitution frequencies higher than the observed values can be easily recovered using our method.

3 RESULTS

Genome-wide mutation pattern

Transitions and Transversions In Figure 1a, we plot the corrected frequencies of the two neighbor-independent transitions A:T → G:C and G:C → A:T (denoted by μ_+ and μ_-, respectively) against the average of the four rather similar transversion frequencies (denoted by ν) for each subfamily of REs collected in the entire genome. From this figure, we observe first that the two neighbor-independent transition frequencies show 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 ($\mu_\pm/\nu$) as the relative transition rates, we find $\mu_+ = (2.74 \pm 0.04)\nu$ and $\mu_- = (5.5 \pm 0.1)\nu$. [An analysis based on only a few families of DNA transposons was already performed by Lander *et al*. (2001), yielding estimates of $\mu_+ \approx 2.5\nu$, and $\mu_- \approx 5\nu$.]

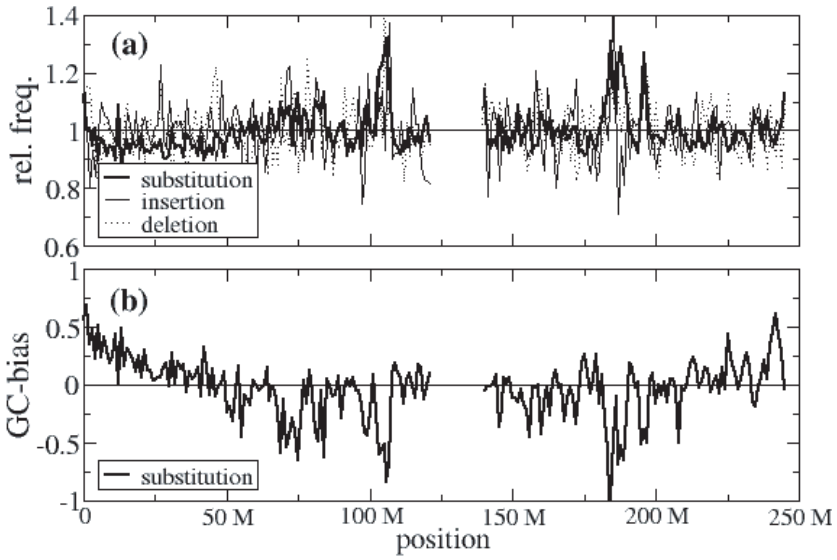

Fig. 2. (**a**) Regional substitution, insertion and deletion frequencies and (**b**) GC-bias of substitutions along human Chromosome 1.

Given the linearity of the data in Figure 1a, it is convenient to use the horizontal axis as the 'time' axis. Calibrating the timescale using estimates of the absolute insertion time of the different Alu subfamilies (Kapitonov and Jurka, 1996), we find each unit of $\nu = 0.01$ in the average transversion frequency to correspond to ∼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 (Kumar and Hedges, 1998; Easteal, 1999; Murphy *et al*., 2001).

CpG-based transition The corrected CpG-based transition frequencies, ω, shown in Figure 1b, present a big surprise: the data clearly present two regimes characterized by very different slopes. This finding is obtained by analyzing REs across different families of SINEs and LINEs, and we verified that it is not an artifact of the ML analysis (Arndt *et al*., 2003). To quantify the extent to which the transition rates changed in the past, we divided the data into two sets of 'young' and 'old' REs with respect to a threshold value ν_0, with the threshold adjusted such that the sum of the squared residuals of linear regressions to the data in both sets is minimal. This minimum was found for $\nu_0 = 0.025$, with a slope of 39.5 ± 2.6 for the young elements and 8.4 ± 2.5 for the old elements. It is natural to identify the two slopes with the relative rates, ω/ν, before and after ν_0. 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 ∼90 Myr ago, corresponding roughly to the time of the mammalian radiation. This conclusion is corroborated by other independent observation by Arndt *et al*. (2003).

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 δ_+ and δ_-, respectively) are computed for each RE subfamily and plotted against the average transversion frequencies in Figure 1c. One observes 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*. (2002).

Regional mutation patterns

The abundance of repetitive elements allows us to estimate also regional substitution patterns along each chromosome. As mentioned above, regional variations in the frequencies of substitutions have already been found by Hardison *et al*. (2003). Here, we can perform a similar regional analysis for mutations along the human lineage alone by repeating the analysis shown in Figure 1 for the human-specific REs residing in each genomic region. This way, we are able to collect detailed information (i.e. all seven substitution frequencies and indel frequencies) with a resolution of 1 Mb.

In Figure 2a, we show the total regional substitution frequency (relative to the genomic averaged value) since the mammalian radiation along the length of Chromosome 1 for each 1 Mb window. (The observations reported here are typical of those exhibited by the other autosomes.) While the variation is confined to the range of ±10% in most regions,

we see distinct substitutional 'hot spots' localized to regions of only a few Mb in length with over 30% in 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, commonly known as the 'genomic isochores' (Filipski *et al*., 1973; Bernardi, 2000). The origin, timing and implications of the human isochore structure is still controversial, [see for a review Eyre-Walker and Hurst (2001)]. Here, we examined regional substitutional biases, i.e. the tendency for substitutions to be GC-enriching or GC-depleting. We plotted in Figure 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%. Our finding implies that telomeres slowly accumulate G and C nucleotides and becomes GC-rich, as has been observed for warm-blooded vertebrates (Perani *et al*., 2000). However, more detailed analysis indicates that this process is unlikely to be the single cause of the human isochore structure since the expected stationary GC-content is still below the observed GC-content (Arndt *et al*., 2003) and other GC-enriching processes acting on larger scales have to be taken into account as well.

4 OUTLOOK

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 to address an outstanding problem facing comparative genomic analysis, i.e. the difficulty of distinguishing sequence homology due to functional constrains from that due to common evolutionary ancestry.

REFERENCES

Arndt,P.F., Burge,C.B. and Hwa,T. (2002) DNA sequence evolution with neighbor-dependent mutation. *Proceedings of the 6th Annual International Conference on Computational Biology* (*RECOMB2002*), Washington DC. ACM Press, New York, pp. 32–38.

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

Bernardi,G. (2000) Isochores and the evolutionary genomics of vertebrates. *Gene*, **241**, 3–17.

Britten,R.J., Baron,W.F., Stout,D.B. and Davidson,E.H. (1988) Sources and evolution of human *Alu* repeated sequence. *Proc. Natl Acad. Sci., USA*, **85**, 4770–4774.

Easteal,S. (1999) Molecular evidence for the early divergence of placental mammals. *Bioessays*, **21**, 1052–1058; Discussion 1059.

Eyre-Walker,A. and Hurst,L.D. (2001) The evolution of isochores. *Nat. Rev. Genet.*, **2**, 549–555.

Filipski,J., Thiery,J.P. and Bernardi,G. (1973) An analysis of the bovine genome by Cs2SO4-Ag density gradient centrifugation. *J. Mol. Biol.*, **80**, 177–197.

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

Hess,S.T., Blake,J.D. and Blake,R.D. (1994) Wide variations in neighbor-dependent substitution rates. *J. Mol. Biol.*, **236**, 1022–1033.

Jurka,J. (2000) Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet.*, **16**, 418–420.

Jurka,J. and Milosavljevic,A. (1991) Reconstruction and analysis of human *Alu* genes. *J. Mol. Evol.*, **32**, 105–121.

Jurka,J. and Smith,T. (1988) A fundamental division in the *Alu* family of repeated sequences. *Proc. Natl Acad. Sci., USA*, **85**, 4775–4778.

Kapitonov,V. and Jurka,J. (1996) The age of *Alu* subfamilies. *J. Mol. Evol.*, **42**, 59–65.

Kumar,S. and Hedges,S.B. (1998) A molecular timescale for vertebrate evolution. *Nature*, **392**, 917–920.

Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Lio,P. and Goldman,N. (1998) Models of molecular evolution and phylogeny. *Genome Res.*, **8**, 1233–1244.

Murphy,W.J., Eizirik,E., Johnson,W.E., Zhang,Y.P., Ryder,O.A. *et al*. (2001) Molecular phylogenetics and the origins of placental mammals. *Nature*, **409**, 614–618.

Perani,P., Caccio,S., Saccone,S., Andreozzi,L. and Bernardi,G. (2000) Telomeres in warm-blooded vertebrates are composed of GC-rich isochores. *Biochem. Genet.*, **38**, 227–239.

Press,W.H., Teukolsky,S.A., Vetterling,W.T. and Flannery,B.P. (1992) *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, Cambridge.

Schwartz,S., Kent,W.J., Smit,A., Zhang,Z., Baertsch,R., Hardison,R.C., Haussler,D. and Miller,W. (2003) Human–mouse alignments with BLASTZ. *Genome Res.*, **13**, 103–107.

Smit,A.F. and Riggs,A.D. (1995) MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. *Nucleic Acids Res.*, **23**, 98–102.

Smit,A.F., Toth,G., Riggs,A.D. and Jurka,J. (1995) Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. *J. Mol. Biol.*, **246**, 401–417.

Waterston,R.H., Lindblad-Toh,K., Birney,E., Rogers,J., Abril,J.F., Agarwal,P., Agarwala,R., Ainscough,R., Alexandersson,M., An,P. *et al*. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.