Nucleic Acids Research, 1996, Vol. 24, No. 6 995–999

A molecular clock based on the expansion of gene families

Yuri A. Trusov* and Paul H. Dear[1]

Institute of Cytology and Genetics, Siberian Branch of Russian Academy of Sciences, Acad. Lavrentiev Avenue 10, Novosibirsk 630090, Russia and [1]MRC Centre, Hills Road, Cambridge CB2 2QH, UK

Received January 25, 1996; Accepted February 8, 1996

ABSTRACT

There is evidence to suggest that eukaryotic genomes are subject to frequent insertions and deletions of non-coding DNA. This may lead to a gradual increase or decrease in genome size, or to a dynamic equilibrium in which the overall size remains constant. We argue, however, that there is a bias favouring an accumulation of non-coding DNA in the proximity of genes. Such bias causes a progressive change in genome structure regardless of whether the overall genome size increases, decreases or remains constant. We show that this change may serve as a 'molecular clock', supplementing that provided by nucleotide substitution rates.

INTRODUCTION

Eukaryotic genomes vary tremendously in size, not only between genera but between closely-related species or even between individuals of a single species whose coding requirements must be closely similar (1–3). This so-called 'C-value paradox' implies that most eukaryotes carry a considerable burden of non-coding DNA with no direct phenotypic effects (4,5).

Such DNA runs the risk of deletion (6,7). Occasionally, such deletion occurs on a large scale, leading to a reduction in genome size (8,9). In most cases, however, the amount of non-coding DNA increases (4,8,10,11), leading in extreme cases to immense genomes such as those of Latimeria and many chordate and teleost species (10,12). This suggests that genomes are in a state of dynamic equilibrium, undergoing frequent insertions and deletions of non-coding DNA (13).

We argue here, however, that there is a bias favouring the accumulation of non-coding DNA in close proximity to genes. This arises because deletions in such regions are liable to disrupt genes and are hence selected against. Conversely, deletions of non-coding DNA in large intergenic regions are less likely to have adverse phenotypic effects.

In consequence, closely-spaced genes will tend to drift apart over time as intervening insertions outweigh deletions. This will lead towards a state where genes are uniformly spaced throughout the genome, and hence where all regions are equally susceptible to deletions. Such 'normalisation' of intergenic distances will tend to occur regardless of whether the overall genome size increases, decreases or remains constant. We present evidence based on clustered gene-families to show that this is the case, and that the process is sufficiently regular to serve as a molecular clock.

THEORETICAL BASIS FOR THE MODEL

Vertebrate genomes contain a large proportion of apparently functionless DNA, consisting of both unique and repeated elements interspersed amongst genes (6,10). A variety of mechanisms have been proposed to account for the insertion, duplication and deletion of such DNA (10,14–20), and the balance between these processes determines whether the genome as a whole expands, contracts or remains of constant size.

Special selective pressures exist on non-coding DNA in the vicinity of genes (we take a 'gene' to include regulatory sequences lying immediately upstream and downstream of the coding region). In a region densely populated with genes, any deletion of non-functional DNA is liable to take with it part of a gene, and so disrupt gene function. Hence deletions will be selected against, the more so as their size (and hence the likelihood of their including a part of a gene) increases. In contrast, insertions will be selectively neutral unless the *point* of insertion lies within a gene, regardless of the size of the insertion. Even tandem duplications within a gene stand a good chance of being tolerated, as they may duplicate part of the gene but leave a copy of the original sequence intact.

The difference between the selective pressure on insertions and on deletions becomes most acute near the ends of genes. Here, insertions which arise by tandem duplication will frequently leave an intact copy of the gene (for example, abcDEFGHIJ...→ abcDEcDEFGHIJ..., where uppercase letters indicate the gene), whereas deletions are particularly liable to destroy control sequences required for the correct initiation and termination of transcription. (An exception to the selective bias against deletions occurs in the case of transposon-like elements which can excise themselves precisely. Deletion of such an element will simply reverse a prior insertion and is not likely to disrupt gene function.)

Qualitative support for this selective bias against deletions comes from mutations in the human β-globin cluster (21). Of several insertions found here, none had apparent effects on health. Deletions, in contrast, were associated with a variety of thalassaemias. Hence, our model states that dense clusters of genes will tend to expand relative to the genome as a whole, due to a 'ratchet effect' whereby insertions are more likely to be tolerated than deletions. The 'ratchet' has its greatest influence at the boundaries between genes and adjacent non-functional sequences. If continued indefinitely

* To whom correspondence should be addressed

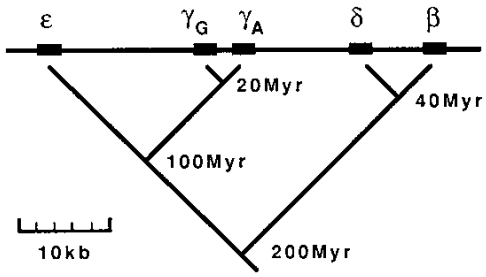

Figure 1. The human β-like globin gene cluster. The physical map (above) is aligned with the evolutionary tree of this gene family (below). The approximate ages of the duplication events are indicated on the tree.

and in the absence of other effects, this would lead to genes becoming uniformly distributed throughout the genome.

Quantitative interpretation of our model is less straightforward. To predict the rate of expansion of a gene cluster accurately, we need to know the frequency with which insertions and deletions of various sizes occur in the absence of selection; these frequencies have not been directly determined, and indirect estimates have been made only in a few special cases (15,16,22–24). However, we can at least make an estimate of the maximum rate at which expansion might be expected to occur. We assume that the initial duplication of a gene gives an intergenic region of ~1 kb (this is supported by examples given below), and that all insertions in this region are selectively neutral whilst all deletions are harmful. The majority of insertions and deletions (other than those involving mobile elements) occur through slippage/mispairing (15,23–25). Based on the data of Di Rienzo *et al*. (22) and Strand *et al*. (25), this process would introduce DNA at an initial rate of ~0.1 kb/Myr. In practice, some insertions will of course be harmful and some deletions selectively neutral, so this rate represents an approximate upper limit on the rate of expansion.

EVIDENCE IN SUPPORT OF THE MODEL

We have sought to test our model by examining the evolution of mammalian gene families. Such families normally arise by tandem duplication of an ancestral gene (often including flanking sequence; 26) and family members usually lie close together when they first arise. Nucleotide substitutions then accumulate in both genes, often followed by further duplications and continued sequence divergence.

According to our model, the accumulation of non-coding DNA between family members should occur progressively, regardless of fluctuations in overall genome size. Hence, physical distances would be greatest between the most distantly-related family members and least between those which, having recently arisen by duplication, show greatest sequence homology. In other words, the evolutionary tree of the gene family (based on sequence divergence) should correspond approximately with its physical map.

A number of factors might conspire to mask this effect. For example, if an ancestral gene 'A' suffered two duplications in rapid succession to yield the gene-family A–B–C, then the distance A–C would obviously be greater than A–B or B–C, despite all three genes having arisen almost simultaneously. Some gene families (for example the Hox clusters) appear to have evolved as functional units in which the spatial relationships of individual genes have probably been selectively preserved (27); such clusters will not expand as predicted by our model. Duplications or inversions involving more than one gene would also confuse the situation; the human immunoglobulin heavy-chain locus has clearly suffered many such events (28), as well as gene-conversions which disguise the apparent ages of duplications. Finally, occasional large insertions or deletions may overwhelm the gradual expansion—due to smaller, more frequent events—which we predict. However, we would expect to see evidence for our model in relatively small gene families which do not appear to have undergone such disruptions. Therefore, we examined those such families for which maps, sequence and, where possible, phylogenetic histories are available.

The human β-globin gene cluster contains five functional genes. Intergenic distances are known precisely and the times of the duplications which gave rise to family members can be inferred phylogenetically (15). These data are summarised as a combined evolutionary tree and physical map in Figure 1; the physical distances between genes correspond well to the estimated times of their divergence. Data for this and other gene families are summarised in Table 1. In the goat and rabbit β-like globin clusters, we find again a correlation between physical distance (29,30) and estimated age since duplication. It should be noted that all of these clusters contain pseudogenes but, for the purposes of our model, we make no distinction between these and any other non-coding sequence. The mouse β-globin-like cluster (comprising genes in the order y–bh0–bh1–b1–b2) fits our model less perfectly. The distances bh0–bh1 and bh1–b1 are as we predict (Table 1), but y–bh0 (2.2 kb) is far smaller than we would expect whilst b1–b2 (14 kb) is larger. The degree of sequence divergence between y and bh0, combined with their orthology to the rabbit b4–b3 and human ε–γ pairs, would lead us to expect an intergenic distance of ~9 kb. We can only assume that a sizeable deletion (~7 kb) has occurred between these genes. Such a deletion must have occurred between the divergence of mouse from rabbit (60 million years ago) and that of mouse from deer–mouse (30 million years ago), as the deer–mouse has a β-like cluster similar to that of mouse (31). In the case of mouse b1–b2, the low degree of sequence divergence leads us to expect an intergenic distance of ~7 kb, as compared with the observed 14 kb. A LINE element insertion accounts for 4.7 kb of the excess. Moreover, the corresponding regions in rat and deer–mouse both contain three genes, of which the outermost members are homologous to the mouse b1–b2 pair (31,32). This strongly indicates that mouse has suffered a recent deletion of a gene between b1 and b2, leaving behind most of the two regions of intergenic sequence.

The human kallikrein gene-cluster comprises three genes (33; Table 1). Although independent estimates of the dates of the duplication events in this cluster are not available from phylogenetic studies, sequence divergences indicate that the pair APS/KLK2 represent the most recent duplication, their common ancestor having diverged from KLK1 much earlier. Again, physical distances between these three genes correspond with the inferred ages of duplications.

The chimpanzee haptoglobin gene cluster consists of three genes in the order Hp–Hpr–Hpp, with intergenic distances of 2.5 and 16 kb respectively (Table 1). The age of the divergence between Hp and Hpr is estimated at 30 million years ago (34), whilst dates for other divergences in the cluster can be inferred only from sequence divergence. Although this cluster fits our model qualitatively (the greatest intergenic distances correspond to the greatest sequence divergence and hence the greatest inferred age since duplication), the distance between Hpr and Hpp (16 kb) is far greater then we would

expect. McEvoy and Maeda (35) found a retroviral insertion of ~8 kb in this region; the remaining 8 kb is approximately in line with our model, though it is slightly larger than expected and may indicate that some additional DNA accompanied the retroviral insertion.

In those cases where the dates of duplication events cannot be determined phylogenetically, we have had to use dates inferred from sequence divergence to test our model. It is therefore important that we should be able to identify any gene-conversion events which may have occurred, as these would greatly reduce such inferred ages. Fortunately, we can recognise gene-conversions in the following way. If gene conversion has occurred between two members of a gene family, then the sequence divergence between them will be less than the divergence between homologues which have existed in related species for comparable lengths of time. That is, gene conversion will lead to apparent intraspecific divergence rates which are far less than interspecific rates. Table 2 lists interspecific divergences between pairs of globin gene homologues in several species, together with the time since the divergence of the species. For most gene pairs, the rate of intraspecific sequence divergence (~0.1%/Myr) is comparable with that of interspecific sequence divergence. The most notable exception is the human γA/γG pair, which show a divergence of only 0.02%/Myr. This supports the suggestion of Slightom *et al*. that this pair has undergone a recent gene conversion (36).

If we can thus recognise gene-conversion events and exclude them from our analysis, we should expect to find a strong correlation between sequence divergence and physical distances between genes, since both are assumed (the latter by our model) to arise from regular, gradual processes. To test this, we considered the pairs of adjacent genes indicated by asterisks in Table 1. Because of the suspected gene conversion between the human γ-genes, we have omitted this pair from our analysis and used in their place the corresponding genes from orangutan, which has a β-like globin cluster of similar structure to human (37). We have also excluded the mouse b1–b2 pair (due to the probable deletion of an intervening gene as described above) and the chimpanzee Hpr–Hpp pair which has suffered a retrovirus-like insertion. We have, however, included the mouse y–bh0 pair, as there is no independent evidence for the large deletion we have postulated. The comparison between sequence divergence and physical distance for these gene-pairs is shown in Figure 2A.

The coefficient of correlation between sequence divergence and physical distance is 0.797 ($z = 3.61$; $0.001 < P < 0.005$). We stress that this strong correlation does not reflect an interdependence of nucleotide substitution and physical separation. Rather, it reflects the fact that these two independent processes are both time-dependent.

Table 1. Sequence divergence, physical distance and approximate age since divergence of some pairs of mammalian genes

		Divergence (%)	Distance (kb)	Age (million years)
Human	ε/γ globin*	11.1 ± 1.9	12	100
	γ_G/γ_A globin	0.35 ± 0.1	3.5	20
	γ/δ globin*	19.0 ± 2.6	14	200
	δ/β globin*	3.3 ± 1.0	5.4	40
	KLK2/APS*	14.1 ± 1.7	12	150
	APS/KLK1*	30.0 ± 2.7	31	310
Orangutan	γ_1/γ_2 globin*	0.9 ± 0.1	3.5	20
Chimpanzee	Hp/Hpr*	4.1 ± 0.7	2.5	30
	Hpr/Hpp	4.8 ± 0.8	16	35
Rabbit	β_1/β_3 globin*	20.2 ± 2.7	14	200
	β_3/β_4 globin*	15 ± 2.4	8	100
Goat	β_c/ε_{II}*	22.2 ± 2.9	13.3	200
	$\varepsilon_I/\varepsilon_{II}$*	14.5 ± 2.4	7	65
Mouse	y/bh_1 globin*	16.5 ± 2.0	2.2	100
	bh_0/bh_1 globin*	3.5 ± 1.1	6.7	45
	bh_1/b_1 globin*	25 ± 2.7	15.9	200
	b_1/b_2 globin	3.3 ± 1.0	14	45

Sequence divergence was calculated according to ref. 38. Asterisks indicate pairs of adjacent genes for which data are plotted in Figure 2. The ages of human globin genes are from ref. 15; those of APS/KLK1/KLK2 were estimated from sequence divergence using the calibrated scale of human globin genes (15). Goat ε and rabbit β-4 genes are homologous to human ε; rabbit β-3 to human γ; goat β-c and rabbit β-1 to human β: the histories of these homologues were therefore assumed to be similar to their human counterparts. Mouse y is homologous to human ε; bh0 and bh1 to human γ; b1 and b2 to human β: the ages of y/bh0 and bh1/b1 duplications are therefore equal to those of their human counterparts. The bh0/bh1 and b1/b2 duplications both exist in mouse and in deer–mouse but not in rabbit. Hence they must both have occurred between 60 (divergence of Murinae and Lagomorpha) and 30 million years ago (divergence Muridae and Cricetidae; 39); we have assumed a value mid-way between these limits. The goat ε-I/ε-II genes correspond to their bovine counterparts, implying that this duplication occurred between 100 (time of mammalian radiation) and 30 million years ago (time of radiation of Bovidae; 39); we have assumed a value mid-way between these limits. The γ-globin gene pairs are orthologous between human, gorilla and orangutan, whilst the lemur and owl monkey have only one γ gene (37); therefore the duplication of the ancestral γ gene occurred between the appearance of the anthropoids (22 million years ago) and the parting of Hominidae from Pongidae (17 million years ago; 39). The age of the Hp/Hpr duplication is from ref. 34; that for Hpr/Hpp is based on data from refs 34 and 35.

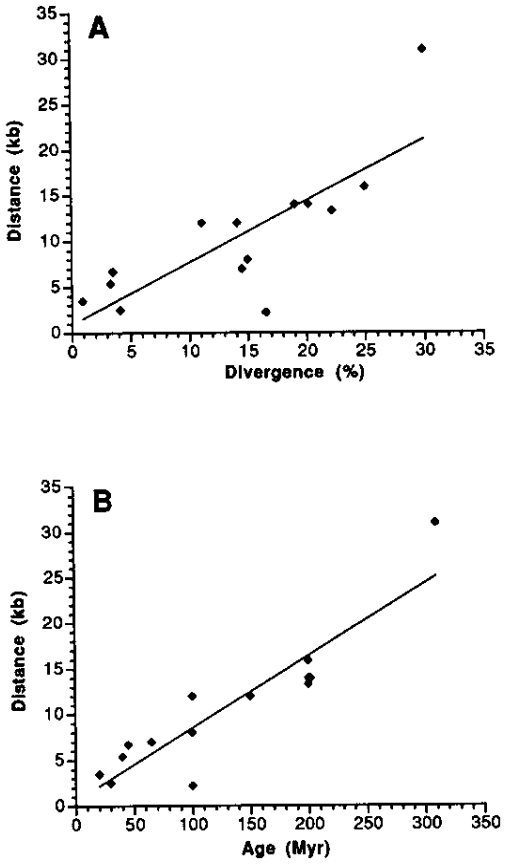

Figure 2. (**A**) Relationship between inter-gene distance and sequence divergence. Only pairs of genes indicated by asterisks in Table 1 were used (see text). (**B**) Relationship between inter-gene distance and age since duplication, using the same gene-pairs as for (A).

A direct estimate of the rate of physical separation of genes proposed by our model is complicated by the fact that exact ages for the relevant duplications are often not known. Where possible, we have used ages based on phylogenetic evidence; where this has not been possible, ages have been based on sequence divergence. Physical distance is compared with age since duplication in Figure 2B. As we would predict, there is a strong correlation ($r = 0.913$; $z = 5.12$, $P < 0.001$) between distance and age.

DISCUSSION

Figure 2B suggests that gene separation can be used as a molecular clock of comparable reliability with that based on nucleotide substitution rates, at least for mammalian gene-clusters over periods of a few tens of millions of years. The intercept of the line with the vertical axis (~0.6 kb if we assume a linear relationship) represents the amount of flanking DNA which is typically duplicated along with genes, and the slope of the line (~0.08 kb/Myr) gives the net rate of accumulation of non-coding DNA between genes. Unlike the clock based on nucleotide substitution, that described here is immune to the effects of gene conversion and may be less susceptible to selective pressures on coding sequences. However, it is vulnerable to disruption by duplications or inversions involving more than one gene, to translocations, or to occasional large insertions or deletions which may overwhelm the gradual, average process.

Table 2. Sequence divergence and age of species divergence between homologous pairs of globin genes

	Divergence (%)	Age (million years)
ε human/ε lemur	5.8 ± 1.3	40
δ human/δ tarsier	7.3 ± 1.5	40
ε human/ε goat	5.3 ± 1.3	100
β human/β_{maj} mouse	14.0 ± 2.2	100
β human/β bovine	9.8 ± 1.8	100
ε human/ε chicken	20.6 ± 2.7	200
ε goat/ε chicken	21.7 ± 2.8	200

Although both clocks are error-prone, the fact that they are disrupted by different factors means that comparison between them can draw attention to interesting events. For example, if two genes are widely separated in space but show little sequence divergence, it may be inferred that either a gene conversion has recently occurred between two long-diverged genes, or that some major upheaval (such as a single large intervening insertion) has occurred. Conversely, an anomalously small intergenic distance between genes of highly-diverged sequence would imply that a large deletion has brought together two genes which were formerly widely separated. Amongst the genes which we have considered we have found some exceptions to our model. All but one of these, however, can be accounted for by insertions, deletions or gene conversions for which there is independent supporting evidence. Only one exception (mouse y–bh0 globin) is unaccounted for by independent evidence, and we would suggest that there has been a substantial deletion which may be revealed as more comparative sequence data become available.

Another molecular clock, based on the overall rate of genome expansion in salamanders, has recently been proposed (11). This may be regarded as a special case of our clock: in salamanders the entire genome tends to expand whereas in mammals (and, we would expect, most species) gene families tend to drift apart despite the overall genome size remaining constant.

In conclusion, we feel that the phenomenon of gene family expansion is worth further investigation. The recent increase in sequencing and mapping efforts should make it possible to test our model in a wider variety of circumstances.

ACKNOWLEDGEMENT

We would like to thank Dr V. A. Berdnikov for helpful criticism and discussion of our model.

REFERENCES

1 Kumar, A. and Rai, K.S. (1990) *Theor. Appl. Genet.* **79**, 748–752.
2 Kuriyan, P.N. and Narayan, R.K.J. (1988) *J. Mol. Evol.* **27**, 303–310.
3 Laurie, D.A. and Bennet, M.D. (1985) *Heredity*, **55**, 307–313.
4 Doolittle, W.F. and Sapienza, C. (1980) *Nature* **284**, 601–603.
5 Orgel, L.E. and Crick, F.H.C. (1980) *Nature*, **284**, 604–607.
6 Dover, G., Brown, S., Coen, E., Dallas, J., Strachan, T. and Trick, M. (1982) in Dover G. and Flavell, R.B. (eds) *Genome Evolution*, Academic Press, London.
7 Saitou, N. and Ueda, S. (1994) *Mol. Biol. Evol.* **11**, 509–512.
8 Sessions, S.K. and Larson, A. (1987) *Evolution* **41**, 1239–1251.
9 Bullock, D.G. and Rayburn, A.L. (1991) *Maydica* **36**, 247–250.

10 Berdnikov, V.A. (1990) Osnovnie factory makroevolucii (in Russian), Nauka, Novosibirsk.
11 Martin, C.C. and Gordon, R. (1995) *J. Evol. Biol.* **8**, 339–354.
12 Hinegardner, R. (1968) *Am. Natr.* **102**, 517–523.
13 Flavell, R.B. (1982) in Dover G. and Flavell, R.B. (eds) *Genome Evolution*, Academic Press, London.
14 Smith, G.P. (1976) *Science* **191**, 528–535.
15 Efstratiadis, A., Posakony, J.W., Maniatis, T., Lawn, R.M., O'Connell, C., Spritz, R.A., DeRiel, J.K., Foget, B.G., Weissman, S.M., Slightom, J.L., Blechl, A.E., Smithes, O., Baralle, F.E., Shoulders, C.C. and Proudfoot, N.J. (1980) *Cell* **21**, 653–668.
16 Jessberger, R. and Berg, P. (1991) *Mol. Cell Biol.* **11**, 445–457.
17 Brown, S.D.M. and Dover, G.A. (1979) *Nucleic Acids Res.* **6**, 2423–2434.
18 Christie, N.T. and Skinner, D.M. (1980) *Nucleic Acids Res.* **8**, 279–298.
19 Donehower, L., Furlong, C., Gillispie, D. and Kurnit, D. (1980) *Proc. Natl Acad. Sci. USA* **77**, 2129–2133.
20 Charlesworth, B., Langley, C.H. and Wolfgang, S. (1986) *Genetics* **112**, 947–962.
21 Weatherall, D.I. and Clegg, J.B. (1982) *Cell* **29**, 7–9.
22 Di Rienzo,A., Peterson, A.C., Garza, J.C., Valdes, A.M., Slatkin, M. and Freimer, N.B. (1994) *Proc. Natl Acad. Sci. USA* **91**, 3166–3170.
23 Tautz, D. and Schlotterer, C. (1994) *Curr. Opin. Genet. Dev.* **4**, 832–837.
24 Dover, G.A. (1989) *Trends Genet.* **5**, 100–102.
25 Strand, M., Prolla, T.A., Liskay, R.M. and Petes, T.D. (1993) *Nature*, **365**, 274–276.
26 Bostok, K.J. and Tyler-Smith, K. (1982) in Dover, G. and Flavell, R.B. (eds) *Genome Evolution*, Academic Press, London.
27 Acampora, D., D'Esposito, M., Faiella, A., Pannese, M., Midlinccio, F.M., Stornaiuolo, A., Nigro, V., Simione, A. and Boncinelli, E. (1989) *Nucleic Acids Res.* **17**, 10385–10402.
28 Cook, G. P., Tomlinson, I.M., Walter, G., Riethman, H., Carter, N.P., Buluwela, L., Winter, G. and Rabbitts, T.H. (1994) *Nature Genet.* **7**, 162–168.
29 Lacy, E., Hardison, R.S., Quon, D. and Maniatis, T. (1979) *Cell* **18**, 1273–1283.
30 Townes, T.M., Fitzgerald, M.C. and Lingrel, J.B. (1984) *Proc. Natl Acad. Sci. USA* **81**, 6589–6593.
31 Padgett, R.W., Loeb, D.D., Snyder, L.R.G., Edgell, M. H. and Hutchison, C.A. (1987) *Mol. Biol. Evol.* **4**, 30–45.
32 Paunesku, T., Stevanovic, M., Radosavljevic, D., Drmanac, R. and Crkvenjakov, R. (1990) *Mol. Biol. Evol.* **7**, 407–422.
33 Riegman, P.H.J., Vlietstra, R.J., Suurmeuer, L., Cleutjens, C.B. and Trapman, J. (1992) *Genomics* **14**, 6–11.
34 Maeda, N. (1985) *J. Biol. Chem.* **260**, 6698–6709.
35 McEvoy, S.M. and Maeda, N. (1988) *J. Biol. Chem.* **263**, 15740–15747.
36 Slightom, J.L., Blechi, A.E. and Smithies, O. (1980) *Cell* **21**, 627–638.
37 Barrie, P.A., Jeffreys, A.J. and Scott, A.F. (1981) *J. Mol. Biol.* **149**, 319–336.
38 Li, W.H., Wu, C.I. and Luo, C.C. (1985) *Mol. Biol. Evol.* **2**, 150–174.
39 Carroll, R.L. (1988) *Vertebrate Paleontology and Evolution.* Freeman and Co., New York.