ELSEVIER

Gene 205 (1997) 191–202

GENE
AN INTERNATIONAL JOURNAL ON
GENES AND GENOMES

Patterns and rates of indel evolution in processed pseudogenes from humans and murids

Ron Ophir, Dan Graur *

Department of Zoology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel

Accepted 20 June 1997

Abstract

Patterns and rates of indel (deletions and insertions) evolution were characterized in 156 independently derived processed pseudogenes from humans and murids (mice and rats). A total of 441 deletions and 161 insertions were unambiguously identified. On a subset of 109 pseudogenes, we verified and confirmed the assumption that indels occur almost exclusively in the pseudogene and, therefore, in comparisons between pseudogenes and their functional paralogs, it is possible to assign polarity to the indel event. By comparing the characteristics of terminal truncations with those of internal deletions, we find support for the hypothesis that truncations are generated through a different pathway than internal deletions. The number of deletions and insertions per pseudogene was found to increase monotonically with time. Deletions occur on average once every 40 nucleotide substitutions, whereas insertions are much rarer, occurring once every 100 substitutions, indicating that the mechanisms involved in deletion formation are most probably different from those responsible for the formation of insertions. The age of the pseudogene, however, explained only 20 and 13%, respectively, of the variation in the number of deletions and insertions per site, indicating that factors other than evolutionary time may play a significant role in the evolutionary dynamics of indel accumulation. Since the rate of substitution has been previously shown to be higher in murids than in humans, we deduce that deletions and insertions accumulate proportionally faster in murids than in humans. Deletions and insertions in murid and human genomes do not contribute significantly to genome size. © 1997 Elsevier Science B.V.

Keywords: Indel evolution; Processed pseudogenes; Retropseudogenes; Truncation

1. Introduction

Processed pseudogenes, or retropseudogenes, arise through the integration of reverse-transcribed mature mRNA molecules into the genome (Vanin, 1985). Because, with very few exceptions, retropseudogenes are non-functional from the moment that they are incorporated into the genome, one may assume that all mutations occurring in them are selectively neutral and may be randomly fixed in populations. That is, the observed patterns and rates of evolution observed in retropseudogenes are a faithful representation of the mutational input. Thus, processed pseudogenes may be used to infer spontaneous mutation patterns and rates.

Whereas the rates, patterns and mechanisms of point-nucleotide mutations have been studied extensively (e.g., Razin and Riggs, 1980; Gojobori et al., 1982), not much is known about insertion and deletions (indels). Previous studies suggested that deletions may be more frequent than insertions (de Jong and Ryden, 1981; Graur et al., 1989; Gu and Li, 1995). However, it is not yet clear whether the two types of indel arise through different mechanisms or whether they are complementary phenomena, i.e. whether or not their formation may be explained by a single molecular model, as attempted by the slipped-strand mispairing model (Levinson and Gutman, 1987).

In this study, we use 156 processed pseudogenes from humans and murids (mice and rats) with a total of 441 deletions and 161 insertions to characterize some features associated with the evolutionary dynamics of deletions and insertions. Specifically, (1) we compared the rates of deletion and insertion for the two taxa, (2) assessed whether the dynamics of insertion is the same as that for deletions, and (3) inferred the contribution of indels occurring in pseudogenes to the genome size.

* Corresponding author. Fax: +972 3 6409403;
e-mail: graur@post.tau.ac.il

2. Materials and methods

2.1. Data

We compiled a database of alignments of 229 processed pseudogenes with their homologous functional mRNA sequences. The pseudogene sequences were collected from the EMBL databank and aligned by using the CLUSTAL W program (Thompson et al., 1994). In case more than one paralogous pseudogene was found for a given functional gene, one should consider the possibility that several pseudogenes may have been derived from one another via duplication after the incorporation event. To ensure the independence of each sample, we constructed a phylogenetic tree by using the neighbor-joining method (Saitou and Nei, 1987), and used an orthologous functional gene from a different species to root the tree. If all the pseudogenes of a functional gene are inferred to have diverged directly from the functional gene (Fig. 1a), they were included in the sample. If some of them are inferred to have diverged from another pseudogene (Fig. 1b), only one of them was included in the sample. The number of independently derived pseudogenes was 156 (93 from humans and 63 from murids, Table 1).

2.2. Age of the pseudogenes

The degree of divergence (or evolutionary distance) between a processed pseudogene and its functional homologue was used as an indication of the age of the pseudogene. The evolutionary distance was calculated by using Kimura's two-parameter model (Kimura, 1980), as well as the numbers of transitions and transversions per site between the two sequences. In all calculations, the length of the processed pseudogene at the time of its incorporation into the genome was assumed to be equal to the length of the coding region in the functional gene minus the terminal truncations in the pseudogene.

2.3. Assessment of indel polarity

Indels are likely to be deleterious in coding regions, so we assigned directionality (or polarity) to an indel event by assuming that the observed gaps are due to events occurring in the processed pseudogenes. In other words, whenever a gap in the alignment appeared in the functional gene, an insertion was inferred, whereas when the gap appeared in the processed pseudogene, a deletion was inferred. To verify the assumption that indels indeed occur almost exclusively in the pseudogene, we inferred the polarity of indels by using the method of Gojobori et al. (1982) on a subset of 109 alignments (69 from humans and 40 from murids) for which an orthologous functional gene from a different species was available.

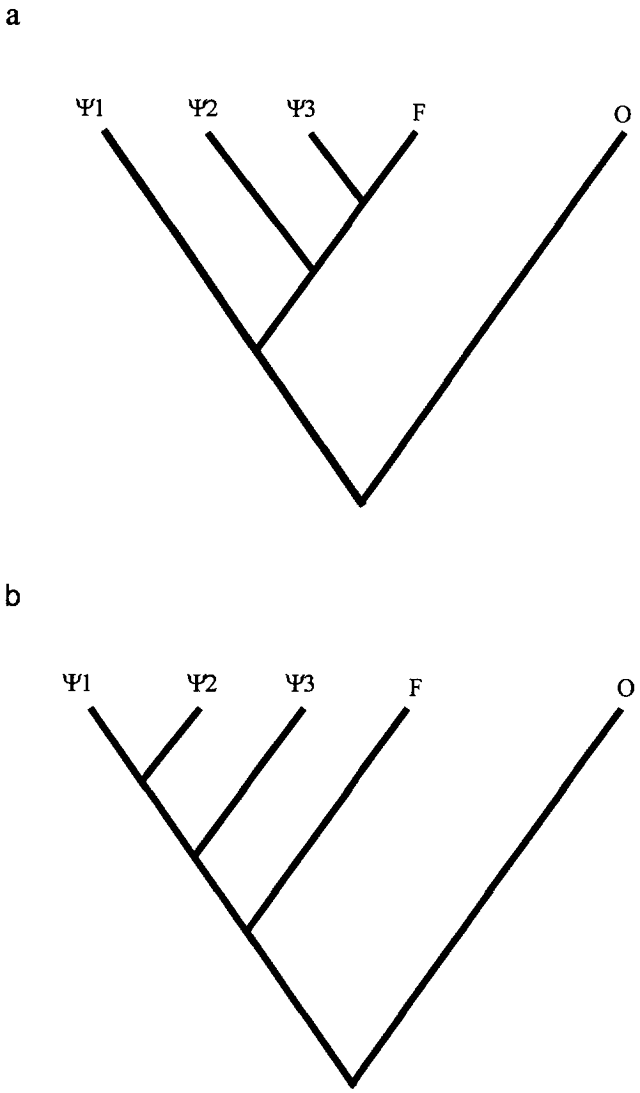

Fig. 1. Schematic representation of divergence of pseudogenes. (**a**) Independent divergence of three pseudogenes (ψ1, ψ2, ψ3) from a paralogous functional gene (F). (**b**) Only one pseudogene diverged from the functional ortholog, whereas the other two pseudogenes derived from the first pseudogene. The trees are rooted by an orthologous functional gene from an outgroup taxon (O).

2.4. Indel characterization

The variables used in the analyses were (1) number of indels per site, and (2) total length of indels as a fraction of the size of the pseudogene at the time of its incorporation into the genome. Since we are interested in mutational events that occurred after the incorporation of the processed pseudogenes into the genome, terminal gaps were excluded from the main body of the analysis (see below).

2.5. Truncations

Some of the processed pseudogenes are truncated at their 3′- or 5′-ends. These terminal gaps are assumed to

Table 1
List of independently derived processed pseudogenes in human and murids and proportion of unchanged nucleotides (p) between the pseudogenes and their paralogous functional genes

Taxon	Gene	Functional gene accession number	Pseudogene accession number	p
rat	diazepam binding inhibitor	M14201	Z11986	0.98
			Z11987	0.93
			Z11989	0.88
	δ-aminolevulinate dehydratase	X04959	J04764	0.95
	calmodulin 1	X13933	X04271	0.91
	calmodulin 2	M19312	M17068	0.93
	cytochrome-c oxidase VIa	X72757	X72759	0.85
	cytochrome-c oxidase VIb	X14208	X16489	0.85
	cytochrome-c oxidase VIc	M20183	M21678	0.86
	cytochrome c	M20622	K03240	0.98
	cytochrome P450	M18335	M18336	0.77
	N-ras oncogene	X68394	X68396	0.97
	tumor suppressor p53	L12046	L07904-10	0.88
	α–tubulin	J00798	J00799	0.98
	S-adenosylmethionine decarboxylase	M21155	M34463	0.86
	extracellular signal-related kinase 1	M61177	M64302	0.95
	small nuclear ribonucleotide particle associated protein	M29293	X73410	0.96
	glucose-6-phosphate dehydrogenase	X07467	M24284	0.96
	glutathione S-transferase	X02904	M14364	0.94
	kininogen	M16455	M22232	0.91
	mannose-binding protein C	M14103	M14106	0.80
	metalothionein-I	J00750	M11797	1.00
			M11795	0.96
	ornithine aminotransferase	M11842	M55178	0.83

Table 1 (*continued*)

	ornithine decarboxylase	M16982	X13417	0.82
	ornithine decarboxylase antizyme	D11372	D11373	0.95
mouse	lactate dehydrogenase A	M27554	M31035	0.95
	γ-actin	M21495	M10142	0.98
			X13052	0.96
	pregnancy-specific glycoprotein	M83341	M83348	0.76
	j-k recombination sequence protein	X17459	X59130	0.88
	thymidylate synthase	M13019	M30774	0.97
	small-nuclear protein	X62648	X60388	0.96
	serfeit locus 3 protein	M21455-61	M20742	0.91
			M20741	0.96
			M20743	0.96
	thymidine kinase	M68489	X13791	0.97
	α2-globin	V00714	V00715	0.82
	mannose-6-phosphate receptor	X64068	X64069	0.98
	centromeric protein C	U03113	L30105	0.92
	creatine kinase B	M74149	M74148	0.96
	casein kinase II α	U17112	M96173	0.95
	small heat shock protein 25	L07577	L11610	0.99
	α-interferon	M28587	M10076	0.83
	leukosialin	X17018	X17017	0.85
	lymphocyte common antigen	M92933	M14343	0.99
	influenza virus resistance protein	M12279	M21038	1.00
			J03368	0.82
	N-ras oncogene	X13664	X06908	0.91
	nucleolin	X07699	M37985	0.92
	tumor suppresor p53	X00741	X01236	0.96
	proopiomelanocortin	J00612	J00613	0.91
	proliferating cell nuclear antigen	X53068	X57798	0.78
			X57799	0.99
	prolactin receptor	X73372	M22957	1.00
	LLrep3 protein	M20632	M20633	0.99

Table 1 (*continued*)

			M20634	0.98
	sterol carrier protein-2	M62361	M91457	0.92
	mast cell serine protease 1	X68803	X78543	0.93
	seven-in-absentia homolog	Z19579	Z19582	0.93
	t-complex polypeptide 1	M12899	D00851	0.88
	uridine kinase	L31783	L31784	0.94
human	β-actin	X63432	M55014	0.80
	γ-actin	X04098	M55082	0.88
	ADP-ribosylation factor 1	M84326	Z21840	0.90
	ADP-ribosylation factor 4	M36341	M31889	0.94
	α-enolase	M14328	X15277	0.93
	adenylate kinase 3	X60673	X60674	1.00
	aldose reductase	X15414	M84454	0.91
	aldolase A	M11560	M21191	0.90
	arginosuccinate synthetase	X01630	K01845	0.93
			K01846	0.93
	ATP synthase C subunit	X69908	X69909	0.93
	transcription factor BTF3	X53280	M90353	0.89
	heat shock protein 70	Y00371	Y00481	0.95
	calmodulin 1	M19311	X13461	0.95
	calmodulin 2	J04046	X52956	0.80
	ceruloplasmin	J05506	M18058	0.97
	creatine kinase β subunit	M16451	M60806	0.86
	casein kinase II α subunit	M55265	X64692	0.99
	Cu/Zn superoxide-dismutase	J02947	M13266	0.85
			M13268	0.86
	cyclophilin	M60457	M63573	1.00
	steroid 5-a-reductase	M32313	M68887	0.95
	D2-type cyclin	M90813	M91003	0.87
	D3-type cyclin	M90814	M90815	0.88
	connexin 43	M65188	M65189	0.97
	cytochrome b5	M22865	M25765	0.91
			X53941	0.81
	cytochrome c oxydase subunit VIb	X13923	M38259	0.93
	cytochrome c	M22877	M22893	0.97
			M22878	0.92
			M22889	0.91

Table 1 (*continued*)

D5 dopamine receptor	M67439	M75867	0.95
apoferritin L	M11147	X03744	0.86
apoferritin H	M14211-2	J04755	0.96
		X03485	0.95
ferredoxin	M18003	M34787	0.96
ferrochelatase	D00726	X69299	0.85
dihydrofolate reductase	X00855-9	J00146	0.93
		M12903	0.94
glyceraldehyde-3-phosphate dehydrogenase	M33197	X01111	0.96
glutathione peroxidase	Y00483	M93083	0.95
glycerol kinase	L13943	X78713	0.98
GM2-associated protein	M76477	L01440	0.91
histone 2a.1b	L19778	K01889	0.84
glucocerebrosidase	D13286	D13287	0.98
high-mobility group protein 17	M12623	X06353	0.93
heat shock protein p27	X54079	X03901	0.93
interferon-α-II-1	M11003	K03013	0.85
high affinity interleukin-8 receptor	M94582	X65859	0.88
keratin 19	Y00503	M33101	0.91
lactate dehydrogenase A	X02152	X02153	0.91
lactate dehydrogenase B	X13794-801	M60601	0.94
laminin binding protein	J03799	L15458	0.96
		X13712	0.95
lipocortin 2	M14043	M62895	0.93
		M62898	0.98
S-adenosylmethionine decarboxylase	M21154	U02035	0.91
metallothionein-I	K01383	M13073	0.87
		M11399	0.91
metallothionein-II	J00271	J00272	0.97
Na/K-ATPase b subunit	X03747	X17162	0.90
arylamine N-acetyltransferase	X14672	X17060	0.80
small nuclear ribonucleoprotein E	X12466	M65126	0.93
		M36001	0.93

Table 1 (*continued*)

neurotrophin-4	M86528	M86529	0.89
nucleophosmin	M23613	L15316	0.92
		L15318	0.94
		L15319	0.96
poly(ADP ribose) polymerase	M18112	L14752	0.92
cAMP-dependent protein kinase regulatory subunit	M18468	L20252	0.89
prothymosin α	M26708	J04799	0.99
		J04800	0.92
		J04802	0.97
p80-coilin	U06632	L06522	0.94
autosomal phosphoglycerate kinase	X05246	K03019	1.00
X-linked phosphoglycerate kinase	V00572	K03201	0.95
B-raf oncogene	M95712	X65188	0.93
ras-related oncogene	X52987	X12534	0.85
c-Ki-ras oncogene	K03209-10	K01912	0.94
ribosomal protein L23	L13799	U06155	0.93
sphingolipid activator	D00422	M81355	1.00
β-tubulin	J00314	K00841	0.96
		K00840	0.90
		J00315	0.76
transcription elongation factor II	M81601	X75159	0.88
tyrosinase	M27160	D00744	0.98
tax-responsive enhancer	D90209	U03712	0.94
tyrosin kinase 3	X72886	X72887	0.94
triosephosphate isomerase	M10036	K03224	0.92
		K03225	0.92
ubiquitin-fusion p52	X56998	M62405	0.91
ubiquitin	X04803	X04801	0.94
glycosyl phosphatidylinositol anchor	D11466	X77457	0.92

be mainly due to abortive transcription during reverse transcription. We tested this assumption by comparing the evolutionary characteristics of terminal gaps with those of non-terminal ones.

2.6. Temporal dynamics of change in pseudogene length

To investigate the contribution of insertions and deletions to genome size, we calculated the proportional change in pseudogene length relative to its length when it became incorporated into the genome as

$$\Delta I = (L_i - L_d)/L_\Psi \qquad (1)$$

where ΔI is the proportional change in a pseudogene length, L_i is the total length of all its insertions, L_d is the total length of all its deletions, and L_Ψ is the length of the processed pseudogene at the time of incorporation, which is assumed to be the length of the functional gene minus the terminal truncations. A positive ΔI value indicates an increase in length, and a negative ΔI indicates a decrease in pseudogene length.

3. Results and discussion

3.1. Validity of assumptions: inference of indel polarity

In an alignment, two character states are possible at a site: presence or absence of a gap. When comparing a pseudogene, its functional paralog, and an orthologous functional gene from a different species, there are three possible categories of outcome in reference to the sharing of the character state: (1) the character state is shared by the pseudogene and the functional paralog (Figs. 2a, b); (2) the character state is shared by the two functional genes (Figs. 2c, d); and (3) the character state is shared by the pseudogene and the functional gene from a different species (Figs. 2e, f); each of these categories includes two cases: (1) two gaps (Figs. 2a, c, e) or (2) one gap (Figs. 2b, d, f). If a gap is the result of a mutational event that has occurred prior to the emergence of the pseudogene (Figs. 2a, b), then this case will not be represented in our set of alignments, since we have identified the gaps by comparing the pseudogene with its functional paralog. Therefore, this category is irrelevant to our study. In the second category (Figs. 2c, d), the minimum-evolution assumption supports the notion that the indel occurred in the pseudogene. The third category (Figs. 2e, f) is contradictory to our working assumption; it indicates that the mutational event occurred in the functional gene.

In a subset of 109 alignments, for which the three pertinent types of sequence were available for comparison, we identified 402 gaps. Of these, only eight were inferred to have occurred in the functional gene after the divergence of the pseudogene. (As expected, all of these gaps were multiples of 3 bp and, therefore, caused no frameshift in the reading frame.) Clearly, the vast majority of gaps (98.1%) occurred in the pseudogene, and we are, therefore, justified in our assumption concerning the polarity of gaps.

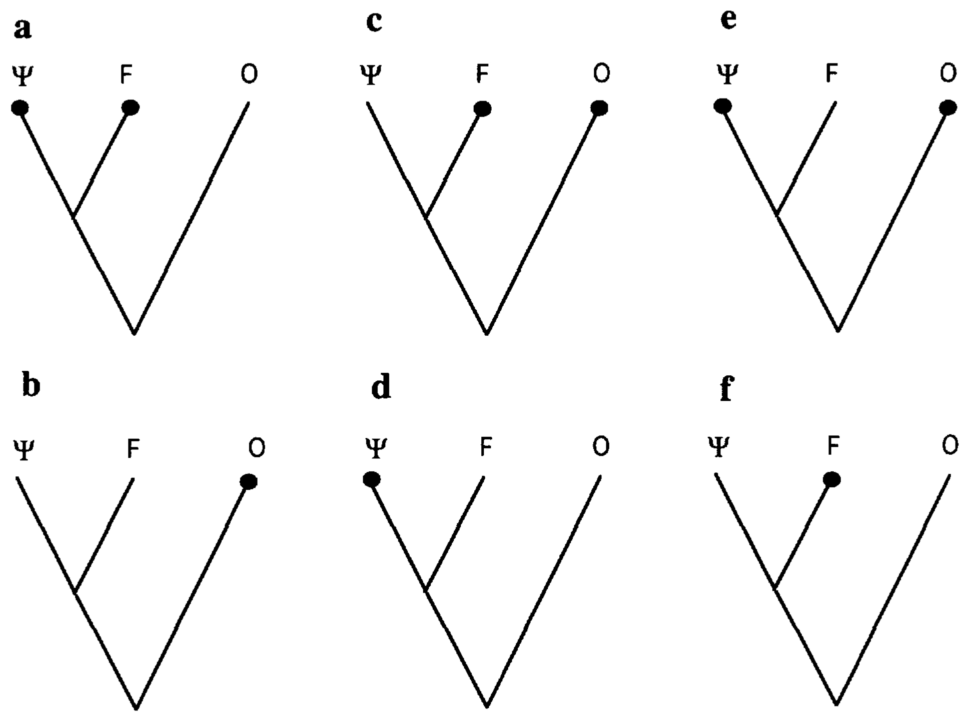

Fig. 2. Possible character-state distribution of gaps (black dots) in comparisons of a pseudogene (ψ), a paralogous functional gene (F), and an orthologous functional gene from a different taxon (O).

Table 2
Distribution of terminal truncation types by taxon and age of pseudogene

Taxon	Age of pseudogene	Not truncated	3′-truncated	5′-truncated	Bilaterally truncated	Total
Human	old	41	10	2	3	56
	young	26	6	5	0	37
	total	67	16	7	3	93
Murid	old	24	9	5	1	39
	young	13	8	2	1	24
	total	37	17	7	2	63

3.2. *Validity of assumptions: truncations*

Some processed pseudogenes are truncated at their 5′- and/or 3′-ends. These 'terminal deletions' are assumed to have been generated through a different pathway than the rest of the deletions. This assumption, however, needs empirical support. To study this problem, we divided the pseudogenes into 'old' and 'young' pseudogenes. The division was based on the mean number of substitutions between the pseudogenes and their functional paralogs. Pseudogenes with distances larger than 0.087 substitutions per site were classified as 'old', and those with distances smaller than 0.087 as 'young.' The numbers of intact pseudogenes, 3′-truncated ones, 5′-truncated ones, and pseudogenes that are truncated at both ends are listed in Table 2.

The distribution of truncations was found to be independent of age in both humans ($\chi^2=4.97$, df=3, $p=0.18$) and murids ($\chi^2=1.11$, df=3, $p=0.78$). These findings strongly support the hypothesis that terminal deletions are generated through a pathway different from that of internal deletions.

The mean sizes of 3′- and 5′-truncations are 301 and 434 nucleotides, respectively. No statistically significant difference was found between the two types ($p=0.51$). Terminal gaps at the 3′-end may be explained by a shift in the initiation of reverse transcription. Terminal gaps at the 5′-end may be explained by an early termination of reverse-transcription. Degradation of the reverse-transcription products may explain truncations at both ends. We found that 33% of all the processed pseudogenes are truncated at one end at least. Contrary to Gu and Li (1995), we find that more pseudogenes are truncated at their 3′-end than at their 5′-end. Since 5′-m7G decapitation is common to all types of degradation (Tuite, 1996), and since the 5′-end of mRNA is the counterpart of the 3′-end of the cDNA, the preponderance of 3′-truncation over 5′-truncation may be explained by mRNA degradation before reverse transcription.

3.3. *Indel characteristics*

The taxonomic distribution of 441 deletions and 161 insertions is shown in Table 3. The mean size of deletions and insertions in murids is 5.91±0.90 and 5.75±1.84, respectively. In humans, the mean size of deletions is 4.67±0.90 and 8.03±2.46, respectively. The mean number of deletions per site is 0.0044±0.0005 in humans and 0.0040±0.0005 in murids. The mean number of insertions per site is 0.0016±0.0003 in humans and 0.0012±0.0001 in murids. The ratio of number of deletions per site to insertions per site is about 3 in both humans and murids. The mean total size of deletions per site is 0.025±0.009 in humans and 0.0031±0.010 in murids. The mean total size of insertions per site is 0.008±0.002 in humans and 0.008±0.002 in murids. The ratio of the total size of deletions per site to insertions per site is about 3–4 in both humans and murids. We found no differences in any of the variables between the two taxa.

Table 3
Numbers of insertions and deletions in human and murid processed pseudogenes

	Murids	Humans	Total
Deletions	197	244	441
Insertions	77	84	161
Total	274	329	603

Therefore, murids and humans exhibit similar indel patterns. The preponderance of deletions over insertions observed in our sample is in agreement with previous studies (de Jong and Ryden, 1981; Graur et al., 1989; Saitou and Ueda, 1994; Gu and Li, 1995).

3.4. *Dynamics of indel evolution*

The dynamics of indel accumulation with the age of the pseudogene is shown in Fig. 3. We found no statistically significant difference in either deletion or insertion accumulation patterns with the age of the pseudogene between humans and murids. However, since murids accumulate nucleotide substitutions twice as fast as primates (Wu and Li, 1985), then the same evolutionary distance in murids represents a time period that is only a quarter of that in primates. Therefore, deletions and insertions accumulate faster in murids than in humans.

Age, however, explains only about 20% of the variation in the number of deletions per site. For insertions,

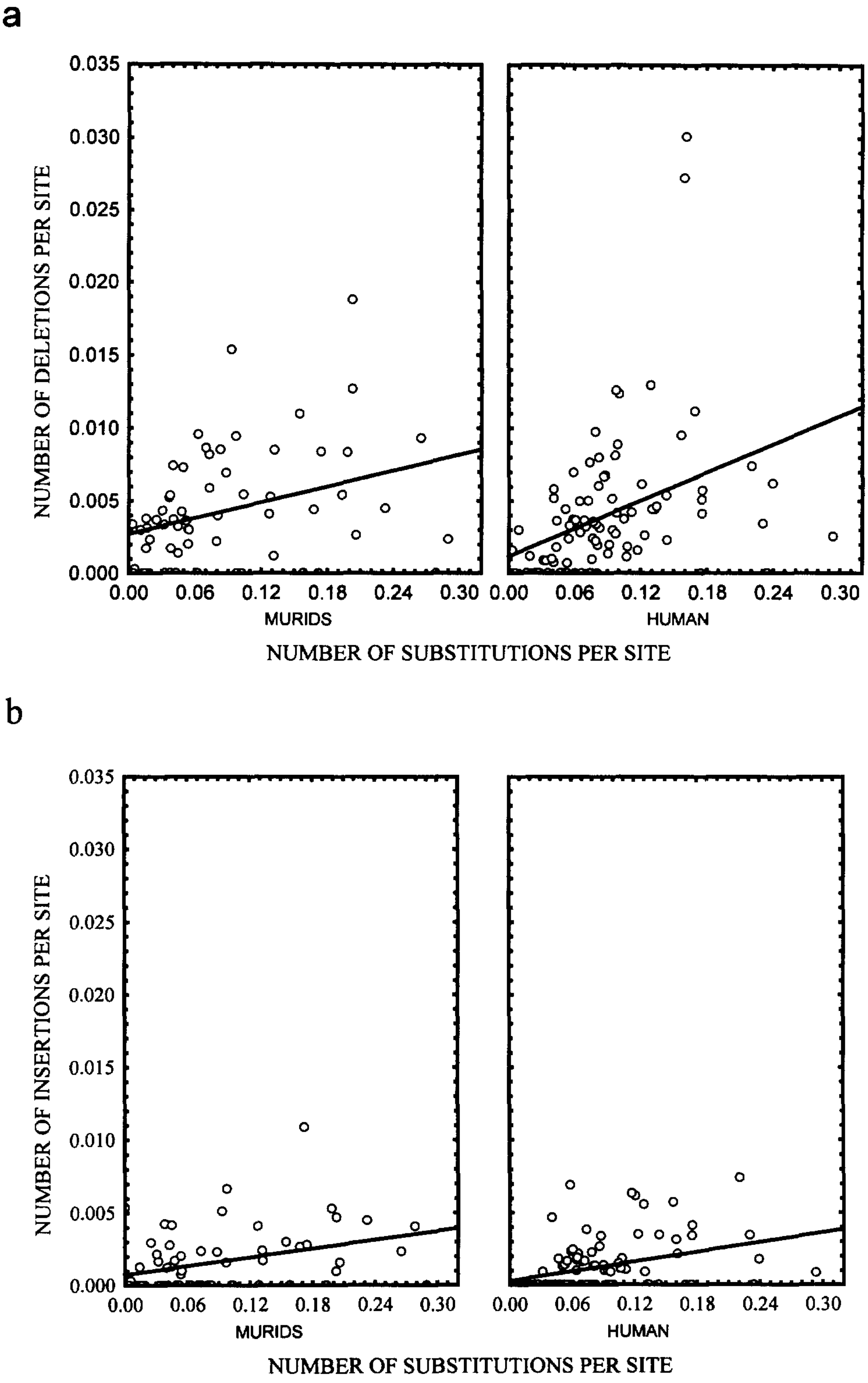

Fig. 3. Number of deletions (**a**) and insertions (**b**) per site as a function of the number of substitution per site between the pseudogene and its functional paralog.

the correlation is 0.36 ($p < 0.00001$), which explains only about 13% of the variation. One reason for the small r values is that we may have sampled two or more indels as one. For example, an inferred 2-bp-long indel may be the result of two 1-bp events occurring at the same site. If this is the case, then our inferring of the number of events would be incorrect. Indeed, we found a significant positive correlation (Spearman $r = 0.22$, $p = 0.01$)

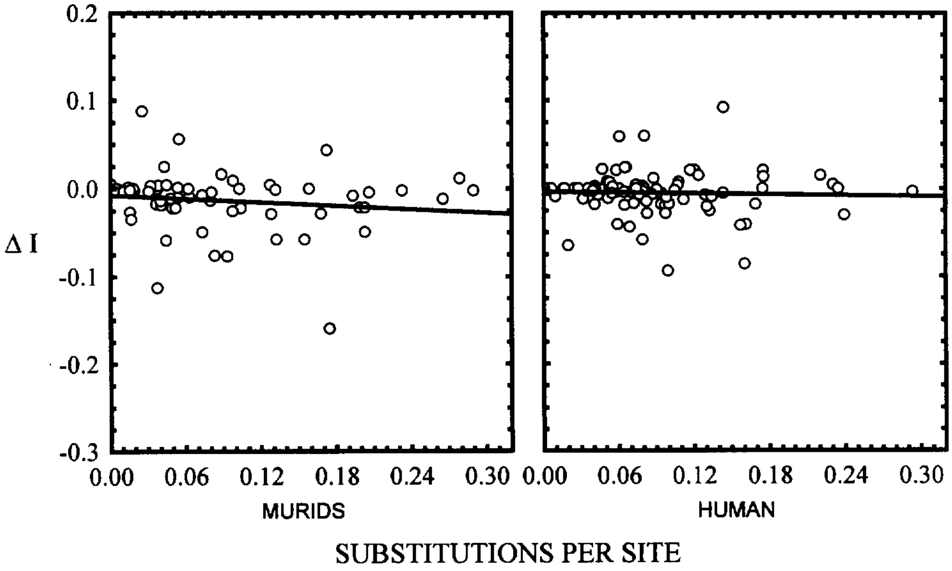

Fig. 4. Changes in pseudogene lengths (ΔI) as a function of pseudogene age as measured by the number of substitution per site between the pseudogene and its functional paralog.

between deletion lengths and evolutionary distance. Insertion lengths are also correlated with the evolutionary distance (Spearman $r=0.20$, $p=0.02$). Notwithstanding, age and clumping of indel events still explain only a small fraction of the variation in the number of indels per pseudogene. Thus, the bulk of the variation must be due to factors other than age. One such factor may be nucleotide composition. This possibility will be explored elsewhere.

From the regression analysis between the number of indels per site and the number of substitutions per site for murids and humans combined, we found that a deletion occurs on average once every 40 nucleotide substitutions, whereas insertions are much rarer, occurring once every 100 substitutions. The fact that we observed different patterns and rates of evolution for insertions and deletions indicates that analyses concerned with indel evolutionary dynamics should be conducted separately for deletions and insertions, as opposed to current practices (e.g. Saitou and Ueda, 1994).

3.5. The contribution of indels to genome size

In both murids and humans, pseudogenes are shorter by approximately 2% than their size at the time of incorporation into the genome, a phenomenon that has been termed 'length abridgment' (Graur et al., 1989). The change in pseudogene size (ΔI) with the age of the pseudogene is shown in Fig. 4. In both taxa, we find a non-significant decrease in pseudogene size with age. Therefore, deletions and insertions do not seem to contribute significantly to genome size. This situation is very different from that in *Drosophila*, where deletions are thought to contribute greatly to genome size evolution (Petrov et al., 1996).

Acknowledgement

We thank Giddy Landan for critical comments. This study was supported by the Magnet Da'at Consortium of the Israel Ministry of Industry and Trade and Silicon Graphics Biomedical.

References

de Jong, W.W., Ryden, L., 1981. Causes of more frequent deletions than insertions in mutations and protein evolution. Nature 290, 157–159.

Gojobori, T., Li, W.-H., Graur, D., 1982. Patterns of nucleotide substitution in pseudogenes and functional genes. J. Mol. Evol. 18, 360–369.

Graur, D., Shuali, Y., Li, W.-H., 1989. Deletions in processed pseudogenes accumulate faster in murids than in humans. J. Mol. Evol. 28, 279–285.

Gu, X., Li, W.-H., 1995. The size distribution of insertions and deletions in human and rodent pseudogenes suggests the logarithmic gap penalty for sequence alignment. J. Mol. Evol. 40, 464–473.

Kimura, M., 1980. A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotides sequences. J. Mol. Evol. 16, 111–120.

Levinson, G., Gutman, G.A., 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol. Biol. Evol. 4, 203–221.

Petrov, D.A., Lozovskaya, E.R., Hartl, D.L., 1996. High intrinsic rate of DNA loss in *Drosophila*. Nature 384, 346–349.

Razin, A., Riggs, D.A., 1980. DNA methylation and gene function. Science 210, 604–610.

Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–425.

Saitou, N., Ueda, S., 1994. Evolutionary rates of insertion and deletion in noncoding nucleotide sequences of primates. Mol. Biol. Evol. 11, 504–512.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680.

Tuite, F.M., 1996. Death by decapitation for mRNA. Nature 382, 577–579.

Vanin, E.F., 1985. Processed pseudogenes: characteristics and evolution. Annu. Rev. Genet. 19, 253–272.

Wu, C.I., Li, W.H., 1985. Evidence for higher rates of nucleotide substitution in murids than in man. Proc. Natl. Acad. Sci. USA 82, 1741–1745.