ELSEVIER

Gene 259 (2000) 81–88

GENE
AN INTERNATIONAL JOURNAL ON
GENES, GENOMES AND EVOLUTION

www.elsevier.com/locate/gene

Densities, length proportions, and other distributional features of repetitive sequences in the human genome estimated from 430 megabases of genomic sequence

Zhenglong Gu, Haidong Wang, Anton Nekrutenko, Wen-Hsiung Li *

Department of Ecology and Evolution, The University of Chicago, Chicago, IL 60637, USA

Received 25 April 2000; received in revised form 4 July 2000; accepted 25 August 2000
Received by T. Gojobori

Abstract

The densities of repetitive elements in the human genome were calculated in each GC content class using non-overlapping windows of 50 kb. The density of *Alu* is two to three times higher in GC-rich regions than in AT-rich regions, while the opposite is true for LINE1. In contrast, LINE2 and other elements, such as DNA transposons, are more uniformly distributed in the genome. The number of *Alu*s in the human genome was estimated to be 1.4 million, higher than previous estimates. About 40% of the autosomes and ~51% of the X and Y chromosomes are occupied by repetitive elements. In total, the human genome is estimated to contain more than 4 million repetitive elements. The GC contents (%) of repetitive elements and their flanking regions were also calculated. The GC contents of almost all kinds of repeats are positively correlated with the window GC contents, suggesting that a repetitive sequence is subject to the same mutation pressure as its surrounding regions, so it tends to have the same GC content as its surrounding regions. This observation supports the regional mutation hypothesis. The only two exceptions are *AluYa* and *AluYb8*, the two youngest *Alu* subfamilies. The GC content of *AluYb8* is negatively correlated with that of its surrounding regions, while *AluYa* shows no correlation, suggesting different insertion patterns for these two young *Alu* subfamilies. This suggestion was supported by the fact that the average genetic distance between members of *AluYb8* in each GC window class is positively correlated with the GC content of the window, but no correlation was found for *AluYa*. *AluYa* is more frequent in Y chromosome than in other chromosomes; the same is true for LTR retroviruses. This pattern might be correlated with the evolutionary history of Y chromosome. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: GC content effect; Regional mutation pressure; Repeat density; Repeat distribution; Y chromosome evolution

1. Introduction

In the past few years our understanding of the human genome organization has increased remarkably. This is especially so with respect to our view of the role of repetitive DNA in the evolution of the human genome. The fact that 14% of the human genome has already been sequenced and annotated provides an excellent opportunity to update our view of the process and patterns of molecular evolution displayed by repetitive elements.

Over 40% of the human genome is made up by four major classes of repetitive elements (Smit, 1999): (1) short interspersed elements (SINEs); (2) long interspersed elements (LINEs); (3) elements possessing long terminal repeats (LTR elements); and (4) DNA transposons.

The most abundant representatives of SINEs are *Alu*s, which are short (~300 bp) elements named after a restriction site they carry. *Alu*s were first derived from 7SL RNA about 80 million years ago (MYA; Kapitonov and Jurka, 1996); however, most *Alu* insertions occurred during the past 65 million years. Based on the presence of diagnostic nucleotide substitutions, *Alu*s are divided into three groups, which are further classified into subfamilies, reflecting the age of individual elements

Abbreviations: ERV, endogenous retrovirus; kb, kilobase pairs; LINE, long interspersed repeat; LTR, long terminal repeat; MaLR, mammalian LTR-retrotransposon; Mb, million base pairs; MIR, mammalian-wide interspersed repeat; MYA, million years ago; SINE, short interspersed repeat.

* Corresponding author. Tel.: +1-773-702-3104;
fax: +1-773-702-9740.
E-mail address: whli@uchicago.edu (W.-H. Li)

0378-1119/00/$ - see front matter © 2000 Elsevier Science B.V. All rights reserved.
PII: S0378-1119(00)00434-0

from oldest (Jo and Jb), to intermediate (Sq, Sp, Sx, Sc, Sg, Sg1), to youngest (Yb8, Ya5, Ya8; Batzer et al., 1996). The *AluJ* group was estimated to have been introduced to the genome 50 to 80 MYA, *AluS* elements were inserted approximately 35 MYA, whereas *AluY* elements are dated back only to 20 MYA (Mighell et al., 1997). Although having a GC level of about 50%, *Alu*s preferentially insert into GC-poor sites (DNA segments having a low relative frequency of guanidine and cytosine nucleotides; Jurka, 1997); however, using mouse/human somatic cell hybrids, Acrot et al. (1995) showed that recent *Alu* insertions appear to be random. The recent analysis by Smit (1999) demonstrated that the majority of *Alu*s (without subdividing into subfamilies) is concentrated in high GC regions, with a GC level of 50–54% (the genome GC average is 42%). Another type of SINE, mammalian-wide interspersed elements (MIRs, ~260 bp), are ancient t-RNA-derived interspersed repeats, which are believed to have spread throughout the genome before the mammalian radiation (Jurka et al., 1995).

LINEs, or non-LTR elements, are long (6–8 kb) GC-poor sequences encoding an endonuclease and a reverse transcriptase (RT) polypeptide. Phylogenetic analysis of RT domain sequences identified 11 distinct LINE groups. In the human genome L1 elements represent the most abundant group of LINEs. The reverse transcriptase encoded by L1s was proposed to be involved in *Alu* transposition (Jurka, 1997).

LTR elements in the human genome have been linked to three classes of human endogenous retroviruses (ERV; Smit, 1999). ELV-L, or mammalian LTR-transposon (MaLR), was inserted into the mammalian genome approximately 70 MYA (Benit et al., 1999), whereas some class I and class II ERVs were introduced to the primate lineage 25–30 MYA (Wilkinson et al., 1994; Andersson et al., 1999).

In contrast to the transposable elements mentioned above, DNA-mediated transposons do not require reverse transcription for the transposition. They propagate themselves through excision and reintegration without an RNA intermediate. They are characterized by terminal inverted repeats not found in other transposons and code for an enzyme, transposase, that catalyzes the excision–reintegration process. The human genome contains at least 14 distinct families of such short (180–1200 bp) degenerate elements that are very ancient and sometimes regarded as transposon fossils (Smit and Riggs, 1996).

In this study, we used the recent data to estimate the distribution pattern of repetitive elements on autosomes and the sex chromosomes. Here we report: (1) the densities and length proportions of repeat elements in genomic regions with different GC levels; (2) the correlation between the GC content of repeats and their surrounding regions; and (3) the average genetic distance between members of *AluYa5* and *AluYb8* subfamilies across different GC levels. We also discuss the possible mechanisms for distinct distributions of different types of repetitive elements.

2. Materials and methods

2.1. Data

The sequences of 2106 non-overlapping contigs (430 Mb, about 14% of the human genome) were downloaded from the Oakridge National Laboratory ftp site (ftp://genome.ornl.gov). Annotations, including description of genes (experimentally defined genes as well as GRAIL and GenScan predictions) and relative coordinates, were retrieved from Oakridge Genome Channel web page (http://genome.ornl.gov).

2.2. Analysis

The Repeat Masker software (with the latest release of the RepBase database update, kindly provided by Dr. Arian Smit) was used to annotate repetitive elements. The masking process was run as four parallel tasks on a Sun Enterprise 4500 server. The *xsmall* option, which masks the repeats with lower-case letters instead of Ns, was used. Each contig was divided into 50 kb non-overlapping windows and the windows were classified into 11 GC content classes based on their GC contents. The density (length proportion and number per 10 kb) of each kind of repetitive element was calculated within each GC content class. Regression analysis was performed between the GC content of a GC class and the GC content of repeats within the class. The average genetic distance calculation was based on Kimura's two-parameter method, using DAMBE (Xia, 2000; version 3.7, kindly provided by Dr. Xuhua Xia, http://web.hku.hk/~xxia). The computer programs for the above analyses were written in C++. The programs and results of analysis will become available at http://ponside.uchicago.edu/~lilab/Gene_Publication/

3. Results

3.1. Repeat distribution over GC levels

Repeat density was studied by dividing sequences into non-overlapping windows of 50 kb. The results of this analysis are shown in Table 1. We considered autosomes and sex chromosomes separately. It was found that 40.7% of autosomes are occupied by interspersed repetitive DNA, while this number is 50.9% and 51.3% in X and Y chromosomes, respectively. The *Alu*s occupy 12.5%, 7.5% and 9.3% of the autosomes, X

Table 1
Densities (number per 10 kb)/proportions (% length) of repeats in each GC content class[a,b]

GC level		<36%	36–38%	38–40%	40–42%	42–44%	44–46%	46–48%	48–50%	50–52%	52–54%	>54%	Total
Size analyzed (Mb)	A[c]	30.6	52.9	60.0	55.3	45.0	37.3	26.6	21.4	15.3	11.8	15.9	372.1
	X	7.3	15.0	15.9	10.2	4.6	2.4	1.4	0.7	0.2	0.2	0.7	58.6
	Y	0.4	0.9	1.8	1.0	0.5	0.3	0.0	0.0	0.0	0.0	0.0	4.9
Alu	A	2.0/5.1	2.8/6.7	3.5/8.7	4.6/11.2	5.5/14.1	6.5/16.5	7.4/18.8	7.9/19.4	8.0/21.2	7.5/21.3	5.0/16.3	4.8/12.5
	X	1.5/4.0	1.9/4.5	2.4/5.6	3.2/10.1	4.7/13.7	6.1/16.4	6.2/17.9	5.6/11.8	5.4/23.7	6.8/21.3	3.5/13.3	2.9/7.5
	Y	2.9/5.6	3.9/9.0	3.8/9.5	3.7/8.7	4.2/12.1	4.9/12.1						3.7/9.3
MIR	A	1.1/1.5	1.3/1.8	1.4/2.1	1.6/2.2	1.8/2.4	2.1/2.8	2.3/3.2	2.7/3.3	2.8/3.3	2.8/3.1	2.3/32.4	1.8/2.4
	X	1.0/1.4	1.3/1.8	1.5/2.2	1.5/2.4	1.7/2.2	1.6/2.1	1.7/2.5	2.2/2.6	2.3/2.9	2.1/2.5	0.9/1.1	1.4/2.0
	Y	0.4/0.6	0.2/0.4	0.5/0.4	0.4/0.4	0.2/0.5	0.3/0.1						0.3/0.4
LINE1	A	3.4/19.5	3.1/19.0	2.9/16.7	2.7/13.5	2.4/10.6	2.1/9.1	1.9/7.2	1.6/6.7	1.5/5.0	1.1/3.6	0.7/3.0	2.4/12.5
	X	5.0/35.3	4.3/33.3	3.5/28.3	2.9/19.0	2.8/14.0	2.4/9.1	1.9/8.5	1.5/7.5	1.1/5.5	1.6/4.0	1.3/5.3	3.6/25.8
	Y	4.5/35.8	4.2/27.8	3.7/25.5	2.8/21.4	1.8/13.1	1.8/13.1	1.5/5.1					3.2/23.5
LINE2	A	1.1/3.0	1.4/3.5	1.5/3.6	1.6/3.6	1.7/3.3	1.7/3.4	1.8/3.8	2.0/3.4	2.0/3.3	2.0/3.3	1.4/2.6	1.6/3.4
	X	1.1/2.6	1.4/3.2	1.5/3.5	1.6/3.7	1.7/3.1	1.7/3.7	1.8/4.1	1.9/5.8	1.8/1.9	1.5/4.4	1.1/2.2	1.5/3.4
	Y	0.4/1.6	0.5/0.5	0.4/0.9	0.3/0.7	0.4/0.6	0.3/0.1						0.4/0.8
LTR_MaLR	A	0.8/3.4	1.0/3.4	1.1/3.5	1.2/3.7	1.3/3.5	1.3/3.2	1.2/3.0	1.1/2.6	0.9/2.1	0.7/1.6	0.5/1.2	1.1/3.2
	X	0.9/4.3	1.2/4.5	1.2/4.1	1.3/3.7	1.2/3.8	1.3/3.3	1.2/3.7	1.1/2.8	0.8/2.6	0.5/2.0	0.5/1.1	1.1/4.0
	Y	0.6/6.3	0.7/1.6	1.0/2.3	0.8/2.6	0.7/2.3	0.7/1.8						
LTR others[d]	A	0.6/3.4	0.8/4.1	1.0/4.0	1.1/4.7	1.4/4.7	1.2/4.0	1.0/3.5	0.0/2.6	0.5/2.1	0.4/1.7	1.0/4.0	
	X	0.9/4.0	1.0/6.1	1.2/6.1	1.5/5.6	1.6/5.8	1.7/5.2	1.6/3.7	0.8/4.9	1.1/0.4	0.3/1.0	0.2/1.6	1.2/5.5
	Y	1.5/11.2	2.0/11.0	2.8/13.6	3.1/15.2	4.2/17.3	2.2/9.6						2.5/13.3
DNA[e]	A	1.3/2.9	1.3/2.9	1.3/3.0	1.4/2.7	1.5/3.0	1.5/2.7	1.6/2.4	1.5/2.4	1.3/1.9	1.1/1.5	0.7/1.2	1.3/2.7
	X	1.0/2.5	1.2/2.7	1.3/2.7	1.4/3.3	1.4/2.6	1.5/2.5	1.3/2.6	1.2/1.8	1.4/1.2	0.5/0.8	0.4/1.0	1.2/2.7
	Y	0.9/2.7	1.0/2.3	0.8/1.5	0.7/1.6	0.7/0.9	0.6/0.9						0.8/1.7
Total	A	10.2/38.8	11.7/41.4	12.7/41.5	14.2/41.8	15.6/41.7	16.5/42.4	17.4/42.4	17.6/41.2	17.1/39.4	15.6/36.6	10.928.2	14.0/40.7
	X	11.3/54.1	12.3/55.9	12.6/52.4	13.5/47.8	15.1/45.2	16.3/42.2	15.7/43.0	14.3/37.1	13.9/38.2	13.3/36.0	8.0/25.6	12.9/50.9
	Y	11.1/63.7	12.6/52.6	13.0/53.6	11.8/50.7	12.2/46.7	10.5/29.5						11.8/51.3

[a] The GC content of 50 kb non-overlapping windows was used to divide the genomic sequences into 11 GC content classes.

[b] For each set of numbers separated by '/', the first is the density (number of repeats per 10 kb) and the second is the length proportion (%) of repeats in the GC content class.

[c] A stands for autosomes, X for X chromosome, and Y for Y chromosome.

[d] 'LTR others' were those identified as LTR elements, but not the LTR_MaLR. Most 'LTR others' are class I and class II endogenous retroviruses.

[e] 'DNA' stands for DNA transposon. The same comment applies to all tables.

chromosome and Y chromosome, respectively. The density of *Alu*s is two to four times higher in GC-rich regions than in AT-rich regions. For example, there are, on average, ≥7.5 *Alu*s per 10 kb for the GC classes of 48–50%, 50–52%, and 52–54%, whereas only two or three *Alu*s per 10 kb for the GC classes of <36% and 36–38% (Table 1). Note, however, that the density of *Alu*s in the GC class of >54% is relatively low (only five per 10 kb), which might reflect the rarity of AT-rich sites in such regions (see later). The density distributions of LINE2 and another kind of SINE, MIR, are similar to that of *Alu*s, but with less dramatic variations. In comparison, the density of LINE1 decreases as the GC content of the window increases. For LTR elements and DNA transposons, the density is highest in medium GC content regions (around 44%). 'LTR others', which consist mostly of class I and class II ERVs, are much more frequent in Y chromosome than in other chromosomes (two to three times). The densities (number per 10 kb) of three youngest *Alu* elements are shown in Table 6. The density of *AluYa* is the highest in Y chromosome, whereas this is not true for *AluYb8* and *AluY*. The length proportions of repetitive elements shown in Table 1 are similar to those estimated by Smit (1999).

3.2. Average length and estimated number of repeats in the human genome

The average lengths of repeats are shown in Table 2. From the comparison between the average and full lengths of repeats we can see that partial repeats are very common in the human genome. In particular, long repetitive elements (LINE1, LTR retrovirus) are truncated more often than shorter elements (SINEs). In Table 3, we estimated the copy number of each kind of

Table 2
Average sizes (bp) of repeats in the human genome[a]

	Autosome	X	Y	Full size[a]
All *Alu*	260.2	263.8	254.0	~300 bp
MIR	133.0	141.1	124.9	~260 bp
LINE1	515.0	711.9	739.5	~6.5 kb[b]
LINE2	218.2	230.6	194.5	~3 kb
LTR_MaLR	302.1	353.7	321.1	1.5–10 kb
LTR others	403.1	464.7	537.9	1.5–10 kb
DNA	203.6	216.8	208.5	80 bp–3 kb

[a] Smit, 1996.
[b] Wilkinson et al., 1994.

repeat in the entire genome. The numbers of *Alu* repeats are estimated to be 1 380 000, 46 400, and 12 700 in the autosomes, X chromosome, and Y chromosome, respectively. Therefore, there may be 1.4 million copies of *Alu* in the human genome. The second most frequent repetitive element is LINE1, whose copy numbers are estimated to be 770 000, 58 800, and 11 000 in the autosomes, X chromosome, and Y chromosome, respectively. According to our estimate, there are more than 4 million repetitive elements in the human genome.

3.3. *Average GC contents of each Alu subfamily and their flanking regions*

The GC content of each *Alu* subfamily and those of 50 bp, 300 bp, and 3 kb flanking regions are shown in Table 4. The GC content of an *Alu*, on average, decreases as its age increases, which is expected because the GC content of an *Alu* tends to decrease with time (the CpG dinucleotide, which tends to evolve fast, is about ninefold more frequent in young *Alu*s than the genome average; Schmid, 1998). The average GC contents of 50 bp flanking regions are all lower than the genome GC average (42%), which is also expected because the exact insertion sites of *Alu*s were found to be generally AT-rich (Jurka, 1997). For longer flanking regions (300 bp, 3 kb), the GC content increases slightly with their length, although we excluded *Alu*s in the flanking regions. However, the average GC content of flanking regions is still lower than 40% in most cases, even if the flanking regions are each 3 kb long, indicating that *Alu* elements actually prefer to insert into sites that are not only AT-rich themselves, but are also surrounded by AT-rich flanking regions. The GC content of flanking regions of *AluYa* is the lowest in almost all columns of flanking regions in Table 4.

Table 4
Average GC contents (%) of *Alu* repeats and their flanking regions

	Repeat	50 bp flanks[a]		300 bp flanks		3 kb flanks	
		+[b]	−[b]	+	−	+	−
AluJb	50.7	36.9	37.0	41.6	39.2	43.6	39.3
AluJo	49.2	37.6	37.5	42.4	39.8	44.2	40.0
AluSq	51.4	37.3	37.5	41.9	39.6	43.9	39.8
AluSp	51.8	37.6	37.0	42.0	39.1	43.8	39.3
AluSx	51.5	37.4	37.7	42.4	40.0	44.0	40.1
AluSg	52.2	37.4	37.1	41.8	39.3	43.6	39.5
AluSc	51.7	37.1	36.7	41.3	38.6	42.9	38.8
AluY	54.0	37.0	36.6	41.2	38.7	43.0	38.8
AluYb8	54.0	37.5	36.0	41.4	38.4	42.6	38.6
AluYa	56.1	36.3	35.3	39.5	36.6	40.8	37.0

[a] The length of each of the two regions flanking the *Alu*.
[b] '+' means including *Alu*s in the flanking regions; '−' means excluding the *Alu*s in the flanking regions.

3.4. *Genetic distances between members of AluYa5 and AluYb8 in each GC content class*

We obtained the sequences of *AluYb8* and *AluYa5* with length >280 bp and calculated their average genetic distances between members within either of these two youngest *Alu* subfamilies in each GC content class. We treated *AluYa5* and *AluYa8* separately to avoid bias in genetic distance calculations. From Table 5 we see that

Table 3
Observed and estimated numbers of repeats in the human genome[a]

Repeat	Observed number of repeats in sequenced regions				Estimated number of repeats in entire chromosomes			
	Autosomes	X	Y	Total	Autosomes	X	Y	Total
Alu	179 380	16 686	1782	197 848	1 379 846	46 350	12 729	1 438 925
MIR	66 532	8321	161	75 014	511 785	23 114	1150	536 049
LINE1	91 043	21 158	1551	113 752	700 331	58 772	11 079	770 182
LINE2	58 354	8493	189	67 036	448 877	23 592	1350	473 819
LTR_MaLR	39 126	6661	381	46 168	300 969	18 503	2721	322 193
LTR others	36 925	6936	1206	45 067	284 038	19 267	8614	311 919
DNA	49 466	7257	392	57 115	380 508	20 158	2800	403 466
Total	520 826	75 512	5662	602 000	4 006 354	209 756	40 443	4 256 552

[a] In this study, about 13%, 36%, and 14% of the autosomes, X chromosome, and Y chromosome, respectively, were analyzed; estimated sizes of each chromosome were from the NCBI homepage.

Table 5
Average genetic distance (d) between members within each GC content class for *AluYb8* and *AluYa5*[a,b]

		<36%	36–38%	38–40%	40–42%	42–44%	44–46%	46–48%	48–50%	50–52%	52–54%
AluYb8	#	33	88	85	78	53	58	42	37	15	15
	d	0.020	0.112	0.074	0.094	0.095	0.132	0.156	0.146	0.129	0.267
		(0.010)	(0.106)	(0.080)	(0.094)	(0.094)	(0.092)	(0.107)	(0.100)	(0.086)	(0.133)
AluYa	#	51	102	98	80	47	44	28	15	10	10
	d	0.018	0.029	0.021	0.044	0.052	0.036	0.043	0.037	0.023	0.018
		(0.008)	(0.027)	(0.010)	(0.067)	(0.054)	(0.039)	(0.043)	(0.022)	(0.014)	(0.007)

[a] In each GC content class, *AluYb8* and *AluYa5* with length >280 bp were used for average genetic distance calculations.

[b] # stands for the number of repeats in each class. The standard error is given in parentheses.

for *AluYb8*, the average genetic distance increases with the GC content of surrounding regions, whereas for *AluYa5*, no correlation exists. Regression analysis was performed and the P values for *AluYb8* and *AluYa5* were 0.016 and 0.523, respectively. Generally speaking, the average genetic distance between members of *AluYb8* is larger than that of *AluYa5*.

3.5. *Relationship between the GC content of repeats and the GC content of their surrounding regions*

Fig. 1 shows the GC contents of each kind of repeat and their surrounding regions. The slope and P value of regression are shown in each chart. There is a significant positive correlation between GC contents of repeats and their surrounding regions, with the exception of *AluYb8* and *AluYa*. For LINE2 and MIR, which are old repeats, their GC contents were nearly homogenized with the surroundings. For other repeats, *Alu*s in particular, the older the repeat is, the sharper the slope. But for *AluYb8*, its GC content is negatively correlated with the GC content of surrounding regions. Another young *Alu* subfamily, *AluYa*, does not show any correlation between its GC content and the GC content of its surrounding regions.

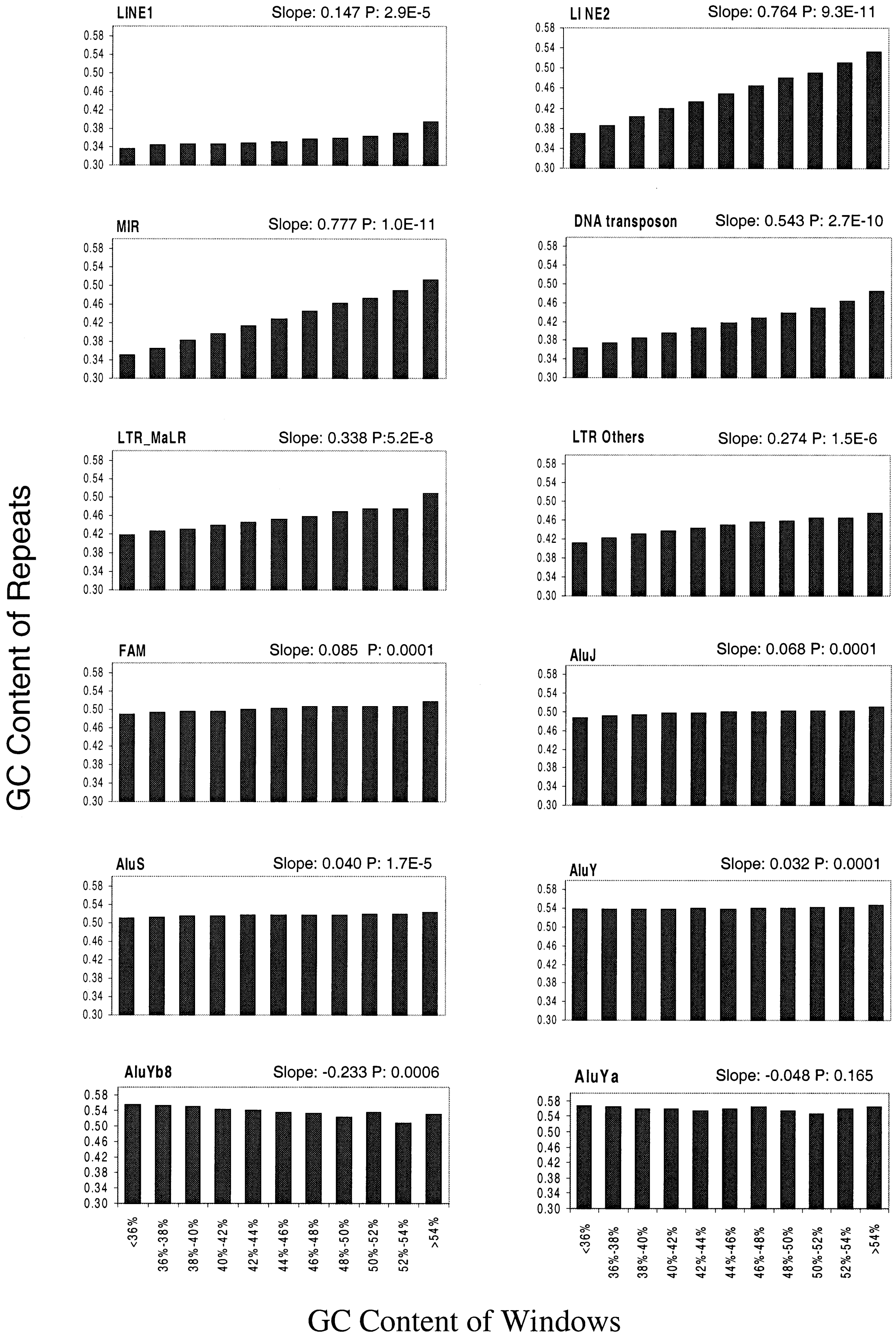
4. Discussion

4.1. *The regional mutation pressure hypothesis*

It has been shown previously that mutation pressure has strong effects on the amino acid composition, species-specific codon preference switches, and some other important evolutionary processes (see Li, 1997). We know that more than 40% of the human genome is composed of repetitive elements, including LINEs, SINEs, LTR elements and DNA transposable elements. These repeats should be subject to the same mutation pressure as their surrounding regions, and because of the general lack of function, their GC contents should tend to be the same as the GC contents of their surrounding regions. To test this, we divided the human genome into 11 GC content classes according to the GC contents of 50 kb non-overlapping windows. The observed positive correlation between the GC content of a class and the GC content of repeats within the class (Fig. 1) supports the regional mutation pressure hypothesis (Wolfe et al., 1989), which postulates that the mutation rate and pattern in a sequence vary with chromosome regions. We observed that the average GC content of each kind of *Alu* repeat is about 2% lower on Y chromosome than on X chromosome. This is interesting, because Y chromosome has a higher mutation rate than X chromosome, leading to a faster decrease in the GC content of *Alu* repeats on Y chromosome. Another observation is that generally the GC contents of repeats are higher in chromosomes with high GC contents than those in chromosomes with low GC contents (data not shown).

4.2. *Insertion patterns of Alu subfamilies*

The young *Alu* subfamilies are good material for studying *Alu* insertion patterns because they have experienced fewer changes than other *Alu* subfamilies. There are two young *Alu* subfamilies, *AluYb8* and *AluYa*. The following differences between these two subfamilies suggest that the insertion patterns of these two subfamilies are different: (1) the average genetic distance between members of *AluYb8* shows a positive correlation with the GC content of its surrounding regions, whereas *AluYa5*, which is the major component of the *AluYa* subfamily, does not (Table 5); (2) the GC content of *AluYb8* shows a negative correlation with the GC content of its surrounding regions, whereas *AluYa* does not (Fig. 1). As the genetic distance between two *Alu* repeats will increase with time, the first observation above suggests that older *AluYb8* members preferred GC-rich regions, whereas younger members preferred AT-rich regions. One argument for this suggestion is as follows: since new *Alu*s are rich in GC, they would be


LINE1
Slope: 0.147 P: 2.9E-5
LINE2
Slope: 0.764 P: 9.3E-11
MIR
Slope: 0.777 P: 1.0E-11
DNA transposon
Slope: 0.543 P: 2.7E-10
LTR_MaLR
Slope: 0.338 P:5.2E-8
LTR Others
Slope: 0.274 P: 1.5E-6
FAM
Slope: 0.085 P: 0.0001
AluJ
Slope: 0.068 P: 0.0001
AluS
Slope: 0.040 P: 1.7E-5
AluY
Slope: 0.032 P: 0.0001
AluYb8
Slope: -0.233 P: 0.0006
AluYa
Slope: -0.048 P: 0.165
0.58
0.54
0.50
0.46
0.42
0.38
0.34
0.30
<36%
36%-38%
38%-40%
40%-42%
42%-44%
44%-46%
46%-48%
48%-50%
50%-52%
52%-54%
>54%
GC Content of Repeats
GC Content of Windows

subject to a weaker mutation pressure if they are located in GC-rich regions than in AT-rich regions. That is, *Alu*s in GC-rich regions are expected to evolve more slowly than those in AT-rich regions. Therefore, the fact that the average genetic distance between members of *AluYb8* is even shorter in AT-rich regions than that of GC-rich regions suggests that *AluYb8* members in AT-rich regions are, on average, younger than members in GC-rich regions. Because the average genetic distance between members of *AluYa5* and the GC content of their surrounding regions do not show any relationship, we might infer that members of *AluYa*, especially *AluYa5*, have no preference for GC or AT-rich regions. This is also supported by the fact that no relationship exists between the GC content of an *AluYa* member and the GC content of its surrounding regions (Fig. 1).

Because *Alu*s use retrotransposase encoded by LINE1 and LINE1 prefers to insert into AT-rich regions, *Alu*s can avoid competition for enzyme with LINE1 elements if they insert into GC-rich regions, which are replicated at different times from AT-regions in the cell cycle. However, their exact insertion sites should be AT-rich because they use retrotransposase encoded by LINE1. This model can explain the slow-down of *Alu* insertion about 30 million years ago and the switch of insertion preference of *AluYb8* found in this paper, because we can imagine that the number of insertion sites decreases with time.

4.3. Y chromosome evolution

Lack of functional constraints in most parts of the Y chromosome and a smaller effective population size of the Y chromosome compared with other chromosomes predict faster evolution of DNA sequences on Y chromosome than on the other chromosomes. However, this would not be true for a Y chromosome region before the development of recombination suppression in the region. In one of their studies on Y chromosome, Lahn and Page (1999) suggested that there were at least four recombination suppression events during the evolution of X and Y chromosomes in mammals, each followed by functional decay in the suppressed region of Y. Based on homologous information in diverse species, they estimated the timing of these four events. The third event happened about 80 to 130 MYA, which was before the insertion of the first *Alu* element and before the insertion of most endogenous retroviruses in the mammalian genome, while the fourth event happened about 30 MYA, when repetitive elements, such as *Alu* and some LTR retroviruses, were very active. Because of the functional decay, the insertion of these repeats before and after this event should have different effects on fitness and thus different fixation rates. In our study, we found that the frequency of *AluYa*, which is very young, is highest in Y chromosome (Table 6). This pattern, i.e. young elements are more frequent in Y chromosome than in other chromosomes, whereas old elements are not, holds also for LTR retroviruses (Table 1; Smit, 1999). However, the densities of two other young *Alu* elements, *AluYb8* and *AluY*, are not the highest in Y chromosome. We can explain this further if we relate it to the *Alu* insertion pattern, because from the discussion in Section 4.2, we speculate that *AluY* and part of *AluYb8* prefer to insert into GC-rich regions, but Y chromosome itself is GC-poor. The GC contents of X and Y chromosome are similar to each other, and from Table 6, we can see that the densities of *AluY* and *AluYb8* are higher in Y chromosome than in X chromosome.

Table 6
Densities (number per 10 kb) of three youngest *Alu* elements

	Autosome	X	Y
AluY	0.553	0.328	0.489
AluYb8	0.016	0.011	0.015
AluYa	0.015	0.010	0.037

5. Uncited reference

Xia (2000).

Acknowledgements

We are grateful to Dr. Arian Smit for providing the current version of the Repeat Masker and for discussions. Extensive help from Dr. Xuhua Xia is appreciated. The data provided by the Oakridge National Laboratory Computational Biology group greatly facilitated this study. Thanks to Richard Blocker for the UNIX system management and Dr. Richard Hudson for generous permission to use his computational facilities. We thank Dr. Satoshi Ota for discussions. This study was supported by NIH grants GM30998 and GM55759.

Fig. 1. Average GC contents of repeats in each GC content class FAM includes FAM, FLAM, and FRAM. *AluJ* includes all the subfamilies in the *AluJ* age group, while *AluS* includes all subfamilies of the *AluS* age group. Regression analysis was performed between the GC contents of repeats and the GC contents of the embedding classes. The middle GC content was used for each class; e.g. 37% was used for the 36–38% class. We omitted the <36% and >54% classes in regression analysis. The results are shown on top of each chart. See Results for description.

References

Acrot, S., Shaikh, T., Kim, J., Bennett, L., Alegria-Hartman, M., Nelson, D., Deininger, P., Batzer, M., 1995. Sequence diversity and chromosomal distribution of 'young' *Alu* repeats. Gene 163, 273–278.

Andersson, M., Lindeskog, M., Medstrand, P., Westley, B., May, F., Mlomberg, J., 1999. Diversity of human endogeneous retrovirus class II-like sequences. J. Gen. Virol. 80, 255–260.

Batzer, M., Deininger, P., Hellmann-Blumberg, U., Jurka, J., Labuda, D., Rubin, C., Schmid, C., Zietkiewicz, E., Zuckerkandl, E., 1996. Standardized nomenclature for *Alu* repeats. J. Mol. Evol. 42, 3–6.

Benit, L., Lallemand, J.B., Casella, J.F., Philippe, H., Heidmann, T., 1999. ERV-L elements: a family of endogeneous retrovirus-like element active throughout the evolution of mammals. J. Virol. 73 (4), 3301–3308.

Jurka, J., 1997. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proc. Natl. Acad. Sci. USA 94, 1872–1877.

Jurka, J., Zietkiewicz, E., Labuda, D., 1995. Ubiquitous mammalian-wide interspersed repeats (MIR) are molecular fossils from the mesozoic era. Nucleic Acids Res. 23 (1), 170–175.

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

Lahn, B., Page, D., 1999. Four evolutionary strata on the human X chromosome. Science 286, 964–967.

Li, W.-H., 1997. Molecular Evolution. Sinauer Associates, Sunderland, MA. chap. 14.

Mighell, A.J., Markham, A.F., Robinson, P.A., 1997. *Alu* sequences. FEBS Lett. 417 (1), 1–5.

Schmid, C.W., 1998. Does SINE evolution preclude Alu function? Nucleic Acids Res. 26 (20), 4541–4550.

Smit, A.F.A., 1996. The origin of interspersed repeats in the human genome. Curr. Opin. Genet. Dev. 6, 743–748.

Smit, A.F.A., 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr. Opin. Genet. Dev. 9, 657–663.

Smit, A.F.A., Riggs, A.D., 1996. Tiggers and other DNA transposon fossils in the human genome. Proc. Natl. Acad. Sci. USA 93, 1443–1448.

Wilkinson, D.A., Mager, D.L., Leong, J.C., 1994. . In: Levy, J.A. (Ed.), Endogenous Human Retroviruses. The Retroviridae vol. 3. Plenum Press, New York, pp. 465–535.

Wolfe, K.H., Sharp, P.M., Li, W.-H., 1989. Mutation rates differ among regions of the mammalian genome. Nature 337, 283–285.

Xia, X., 2000. Data Analysis in Molecular Biology and Evolution. Kluwer Academic, Dordrecht.