doi: 10.1046/j.1529-8817.2005.00163.x

Worldwide Genetic Variation at the 3′-UTR Region of the *LDLR* Gene: Possible Influence of Natural Selection

N. J. R. Fagundes[1,2,*], F. M. Salzano[2], M. A. Batzer[3], P. L. Deininger[4] and S. L. Bonatto[1]

[1] *Centro de Biologia Genômica e Molecular, Faculdade de Biociências, Pontifícia Universidade Católica do Rio Grande do Sul, 90619-900 Porto Alegre, RS, Brazil*

[2] *Departamento de Genética, Universidade Federal do Rio Grande do Sul, 91501-970 Porto Alegre, RS, Brazil*

[3] *Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, USA*

[4] *Tulane Cancer Center, Tulane University Health Center, New Orleans, LA 70112, USA*

Summary

The low density lipoprotein receptor gene (*LDLR*) contains many *Alu* insertions, and is especially *Alu*-rich at its 3′-untranslated region (3′-UTR). Previous studies suggested that the *LDLR* 3′-UTR could regulate gene expression by the stabilization of its mRNA. Given the faster *Alu* evolutionary rate, and wondering about its consequences in a possibly regulatory locus, we have studied ∼800 bp of 222 chromosomes from individuals of African, Asian, Caucasian and Amerind ancestry, to better understand the evolution of the worldwide genetic diversity at this locus. Twenty-one polymorphic sites, distributed in 15 haplotypes, were found. High genetic diversity was observed, concentrated in one *Alu* insertion (*Alu* U), which also shows a fast evolutionary rate. Genetic diversity is similar in all populations except Amerinds, suggesting a bottleneck during the peopling of the American continent. Three haplotype clusters (A, B, C) are distinguished, cluster A being the most recently formed (∼500,000 years ago). No clear geographic structure emerges from the haplotype network, the global F_{st} (0.079) being lower than the average for the human genome. When ancestral population growth is taken into account, neutrality statistics are higher than expected, possibly suggesting the action of balancing selection worldwide.

Keywords: nuclear sequence variation, balancing selection, *Alu* insertions

Introduction

The low density lipoprotein receptor (LDLR) is a cell surface glycoprotein that plays a key role in maintaining normal plasma cholesterol levels, mediating the endocytosis of LDL and other cholesterol-carrying particles. Defects in the receptor pathway lead to familial hypercholesterolemia (FH), a common monogenic disorder in humans (Goldstein *et al*. 1995). The *LDLR* gene, located on chromosome 19p13.1-3, consists of 18 exons spanning 45 kb, and contains multiple *Alu* insertions which are involved in several large deletions causing FH (Lehrman *et al*. 1987; Goldstein *et al*. 1995).

A high *Alu* density occurs at the 3′-untranslated region (3′-UTR) of this gene, whose 2.5 kb comprise almost half of the whole mRNA (5.3 kb) (Yamamoto *et al*. 1984). Several studies showed evidence that the 3′-UTR of the *LDLR* gene can be an important site of post-transcriptional regulation (Goto *et al*. 1997; Wilson *et al*. 1997, 1998; Knouff *et al*. 2001; Nakahara *et al*. 2002), and Wilson *et al*. (1998) also suggested that the *Alu*-rich portion of this region can function as a cytoskeleton-binding domain, playing a direct role in gene regulation by enhancing the mRNA half-life.

In non-human primates, the middle part of this 3′-UTR contains two complete *Alu* insertions (named U for upstream and D for downstream). After the

[*] Corresponding author: Nelson J. R. Fagundes, M.Sc., Centro de Biologia Genômica e Molecular, Faculdade de Biociências – PUCRS, Av Ipiranga 6681, Prédio 12C, sala 172, 90619-900 Porto Alegre, RS, Brazil. Fax +55-51-3320-3612. E-mail: nrosa@pucrs.br

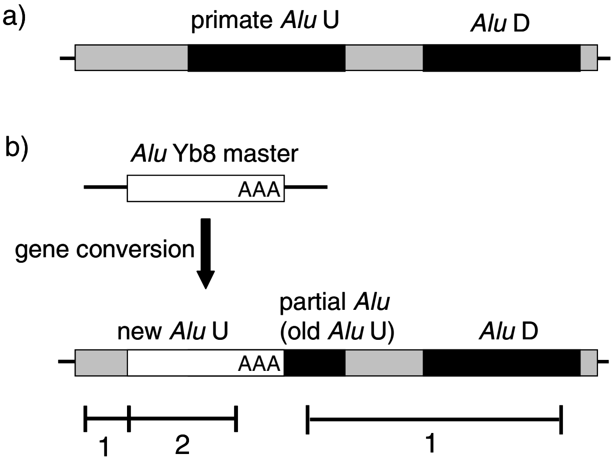

Figure 1 Schematic representation of the arrangement of *Alu* insertions at the 3′-UTR of *LDLR*. a) Genomic organization in non-human primates, containing two elements from an old (Sx) *Alu* subfamily (black boxes). b) In the human lineage, a gene conversion occurred, replacing most of the old, Sx, *Alu* U by a new, human specific, Yb8 subfamily (white box). The new *Alu* U poli A tail is also represented. 1- Homologous human *vs.* chimpanzee sites (non-*Alu* U). 2- *Alu* U sites, for which no homology exists in chimpanzee.

human-chimpanzee divergence, however, *Alu* U was partially replaced by a new insertion from the human-specific Yb8 subfamily through a gene conversion event (Fig. 1). Therefore, in the human lineage, two complete (the new, human-specific, *Alu* U and *Alu* D) and one partial insertion (the old *Alu* U) exist (Kass *et al.* 1995).

The insertion of *Alu* elements in functional loci can modulate gene expression either directly, or by changing the methylation pattern at CpG sites (Batzer & Deininger, 2002). A high CpG content gives *Alu* elements a slightly faster evolutionary rate compared to the rest of the human genome (Chen & Li, 2001). In regulatory sites, such an increased rate can also affect gene function by point mutations (Labuda *et al.* 1995).

To better understand the evolution of the *Alu*-rich region at the 3′-UTR of the *LDLR* gene, with special attention to its worldwide genetic variation, we have studied 222 chromosomes from African, Asian, Caucasian and Amerind samples covering ∼800 bp of this locus. Our data support the idea that the region has an important regulatory function, suggesting that the existing diversity patterns were maintained by balancing natural selection possibly for more than 500,000 years.

Subjects and Methods

Samples

After appropriate informed consent information, DNA from 111 individuals of African, Asian, Caucasian and Native American ancestry were studied. The African or African-derived sample consisted of 25 subjects, namely three Biaka Pygmies, three Mbuti Pygmies, six Nilotic, six Bantu, one Kordofanian, and six African American individuals. The Asian group consisted of 24 persons: four Vietnamese, two Cambodian, five Chinese, three Korean, four Japanese, two Taiwanese, and four Asian individuals of unknown origins. Twenty-six Caucasians were studied, consisting of six German, nine French, one Egyptian, seven Cypriot, one Breton, one Swiss and one Hungarian subjects (these samples had been previously investigated by M.A. Batzer, P.L. Deininger, or L.B. Jorde's groups for other systems). The Native American sample consisted of DNA from 36 individuals, with the following tribal affiliations: seven Aché, six Gavião, six Suruí, five Waiwai, six Xavante, and six Zoró individuals (these samples were collected by F.M. Salzano or his group, and studied for a number of different genetic markers).

PCR Amplification and Sequencing

All samples were amplified using primers F1039 (5′-ACTTCAAAGCCGTGATCGTGA-3′) and R2008 (5′-TGCAACAGTAACACGGCGATT-3′), designed to span 950 bp of the 3′UTR of the *LDLR* gene. Standard amplification conditions were 10-40 ng of genomic DNA, 1.5 mM of $MgCl_2$, 0.2 mM of each dNTP, 0.2 μM of each primer, and 0.5 U of *Taq* DNA polymerase. Cycle conditions were 94°C for 1 min, 60°C for 2 min, and 72°C for 2 min, with an initial denaturing step of 94°C for 1 min and a final extension step of 72°C for 10 min.

PCR products were sequenced on both strands using amplification primers plus PS (5′-ACGGAGTCTCGCTCTGTCGC-3′) and F1501 (5′-ACCATGCATGGTGCATCAGCA-3′) with either BigDye (Applied Biosystems) or ET terminators chemistry (Amersham Biosciences) following manufacturer's protocols, and read in ABI310 (Applied

Biosystems) or MegaBace1000 (Amersham Biosciences) automated systems. All chromatograms were checked by eye to search for possible single nucleotide polymorphisms (SNPs), insertions or deletions (indels). All putative SNP sites were confirmed by exhaustive re-sequencing.

The region sequenced included 784 bp, corresponding to sites 3625 to 4486, with the exception of the *Alu* U A-rich tail (positions 3992 to 4069) (Yamamoto *et al.* 1984).

Haplotype Assignment

Haplotype determination was performed with the PHASE program (Stephens *et al.* 2001a) which implements a Bayesian method of haplotype reconstruction based on genealogies reconstructed from coalescent theory under a Markov Chain Monte Carlo framework. This approach seems to outperform other strategies such as the maximum likelihood expectation maximization algorithm in most cases (Stephens *et al.* 2001a). All searches were conducted with 1,000,000 replications.

Evolutionary Rate

We have partitioned the region into two groups of sites, the first including only the human-specific *Alu* U, and the second including all the other sites (called non-*Alu* U sites – see Fig. 1). For this second group, estimation of the evolutionary rate is straightforward, by calculating the average Kimura-2P distance between human and chimpanzee sequences, and using a calibration point of six million years (Myr). The Kimura-2P model was selected by a maximum likelihood approach (Posada & Crandall, 1998).

For the estimation of the human *Alu* U evolutionary rate, we first estimated the age of insertion of this element in the human genome. To this end, we calculated the mean sequence divergence between the different *Alu* U haplotypes and the *Alu* Yb8 master sequence (Batzer *et al.* 1996) using only the slowly evolving conserved sites (termed CONSBI, for conserved before insertion), for which an evolutionary rate of 0.16% per Myr was estimated (Britten, 1994) using Kimura-2P distance. We then used the average Kimura-2P distance between *Alu* U haplotypes and Yb8 master sequence, considering all positions and the previously calculated insertion age, to estimate an evolutionary rate for the whole *Alu* U region. All distance estimations were performed in the Mega2 program (Kumar *et al.* 2001) and standard errors were calculated by 1,000 bootstrap replications.

Data Analysis

Summary diversity statistics such as the nucleotide (π) and haplotype diversity (H) indices (Nei & Kumar, 2000), as well as the population scaled mutation parameter (θ_w) (Watterson, 1975; Nei & Kumar, 2000), together with their standard errors, were calculated in the DnaSP 3.53 package (Rozas & Rozas, 1999), which was also used to calculate Tajima's (Tajima, 1989) and Fu and Li's (Fu & Li, 1993; Fu, 1997) neutrality tests and their statistical significance.

Since the standard statistical significance for neutrality tests is based on a stationary population, and there are several lines of evidence suggesting that the human population has grown dramatically over the past 100,000 years (Ruhlen, 1994; Stiner *et al.* 1999; Excoffier, 2002), we have also tested for the effect of such growth on the values of the neutrality statistics using the method of Wooding *et al.* (2004) with the DFSC 1.0 program (http://www.xmission.com/∼wooding/DFSC/). We have thus iteratively simulated the value of these statistics under different demographic scenarios to cover a demographic expansion by a factor of 1-fold (no-growth) to 1,000-fold, beginning 0 to 200,000 years ago, using theta and the number of segregating sites observed in the sample as input parameters.

Analysis of molecular variance (Excoffier *et al.* 1992), F_{st} (Wright, 1969), and a test of genetic differentiation based on haplotype frequencies (Raymond & Rousset, 1995) were estimated using Arlequin 2.0 (Schneider *et al.* 2000). A haplotype network was derived using the median-joining algorithm (Bandelt *et al.* 1999) and the Network 3.1.1.1 software (www.fluxus-engineering.com). The network was rooted by the inclusion of a chimerical haplotype consisting of the chimpanzee sequence for the non-*Alu* U sites merged with the *Alu* Yb8 master sequence.

To obtain the time to most recent common ancestor (T_{MRCA}) of the sample, as well as to infer the time of each mutation in the genealogy, we used the coalescent approach of Griffiths & Tavare (1994) implemented in the Genetree 9.0 program (http://www.stats.ox.ac.uk/∼griff/software.html), which uses a Markov chain to simulate a sample of genealogies. All searches we conducted by running 100,000 replications.

Results

Haplotype Assignment

We have identified 21 polymorphic sites within the 784 bp sequenced in the 222 chromosomes, considering the indel at positions 3,818-3,820 as a single polymorphism. On the other hand, the two adjacent polymorphisms at positions 3800 and 3801 were considered independently, as they consist of a deletion and a nucleotide substitution, respectively, and likely have independent origins. Of the 111 individuals tested, 42 had their haplotypes experimentally determined, since they were either homozygous or heterozygous for a single polymorphic site. The remaining individuals had their haplotype assignment estimated by PHASE. When pooling all individuals into a single population, the analysis indicated the existence of 16 haplotypes, which are shown in Table 1 along with their frequencies among continents.

Haplotypes were split into three clusters, named A, B and C according to the presence of specific deletions at positions 3818-3820 for cluster A and 3800 for cluster C (see also the *Haplotype Network* section below). All haplotypes that are present in more than 20 chromosomes are found in all continental groups. The exception is haplotype C1-1 which was not found in Asians. However, this exception could be due to a sampling artifact, since this haplotype is relatively frequent among Amerinds and should have reached the Americas through Asia. Alternatively, it could have been present in the ancestral (and polymorphic) population of Asians and Americans, being maintained in America but not in Asia after the split of these populations. It is also of interest that Native Americans did not present any haplotype that was not observed in the other continents.

All the sequences of the inferred haplotypes were submitted to GenBank and are available through accession numbers AF381773-AF381976 and AY496300-AY496419.

Evolutionary Rates

For the 502 non-*Alu* U sites, an evolutionary rate of 0.166% per Myr was estimated using the average Kimura-2P distance between human and chimpanzee haplotypes. This rate is equal to that suggested by Chen & Li (2001) as the average *Alu* insertion evolutionary rate, and it is worth noting that ∼82% of the non-*Alu* U sites consist of the complete *Alu* D and the partial older *Alu* U element.

For *Alu* U, 193 CONSBI sites were examined, furnishing an estimated age of 3.26 Myr for its insertion at the *LDLR* locus, in good agreement with the 3.42 Myr suggested by Carroll *et al.* (2001) for the origin of the Yb8 subfamily. Thus, the use of the estimated age of insertion is unlikely to overestimate the evolutionary rate, since we would not expect any Yb8 element to be older than 3.42 Myr. Considering all 285 *Alu* U sites and the estimated age of 3.26 Myr, a rate of 0.632% per Myr is obtained. This estimate is significantly higher than the average *Alu* mutation rate of 0.166% per Myr discussed above (Chen & Li, 2001). The reasons for such a high evolutionary rate are unclear. While *Alu* U may simply represent a neutral outlier with respect to its evolutionary rate, it is also possible that other evolutionary processes, such as natural selection or gene conversion, may have played an important role in its evolution.

Genetic Diversity

Summary genetic diversity statistics for each continent and for the whole sample are given in Table 2. Considering all data, the *LDLR* 3′-UTR region displays a considerable amount of genetic diversity, the mean value for π being 0.522% ± 0.010%, considerably higher than other highly variable human autosomal loci, such as the *PON1* promoter ($\pi = 0.190\% \pm 0.008\%$, Koda *et al.* 2004), *GYPA* ($\pi = 0.300\%$, Baum *et al.* 2002), *CCR5* ($\pi = 0.290\% \pm 0.170\%$, Bamshad *et al.* 2002), *LPL* ($\pi = 0.200\%$, Clark *et al.* 1998), and β-globin ($\pi = 0.180\%$, Harding *et al.* 1997), as well as for *X*-linked loci such as *Dmd44* ($\pi = 0.141\% \pm 0.079\%$, Nachman & Crowell 2000), and *PDHA1* ($\pi = 0.180\%$, Harris & Hey 1999), and also large-scale SNP typing studies ($\pi = 0.075\%$, [Sachidanandam *et al.* 2001], and $\pi = 0.058\%$ [Stephens *et al.* 2001b]). The studied region of

Table 1 Inferred haplotypes for the 3′UTR *LDLR* locus and their frequencies among continents

																					Haplotype count				
Position	**3709**	**3751**	**3800**	**3801**	**3812**	**3818**	**3840**	**3846**	**3852**	**3873**	**3876**	**3920**	**3924**	**3938**	**3960**	**3977**	4095	4135	4185	4328	4473	Afr	Asn	Cauc	Ame	World
Ref	c	g	a	g	g	act	c	c	t	a	c	c	c	t	t	c	t	a	g	t	c					
A1-3	c	a	a	g	g	—	c	c	c	a	c	g	c	t	c	c	c	a	t	c	c	7	10	17	27	61
A1-4	c	a	a	g	g	—	c	c	c	a	c	g	c	t	c	c	c	g	t	c	c	0	1	1	0	2
A1-5	c	a	a	g	g	—	c	c	c	a	c	g	c	t	c	c	c	a	t	c	t	1	0	0	0	1
A2-3	c	a	a	g	g	—	c	c	c	a	c	g	a	t	c	c	c	a	t	c	c	0	5	0	5	10
A3-3	c	a	a	g	g	—	c	c	c	g	c	g	c	t	c	c	c	a	t	c	c	0	1	0	0	1
A4-3	t	a	a	g	g	—	c	c	c	a	c	g	c	t	c	c	c	a	t	c	c	2	0	0	0	2
A5-3	c	a	a	g	g	—	c	c	c	a	c	g	a	t	c	t	c	a	t	c	c	0	1	0	0	1
B1-1	c	g	a	g	g	act	c	c	t	a	c	c	c	t	t	c	t	a	g	c	c	7	4	1	8	20
B1-2	c	g	a	g	g	act	c	c	t	a	c	c	c	t	t	c	t	a	g	t	c	2	5	22	19	48
B2-2	c	g	a	g	a	act	c	c	t	a	c	c	c	t	t	c	t	a	g	t	c	0	0	1	0	1
B3-2	c	g	a	g	g	act	t	c	t	a	c	c	c	t	t	c	t	a	g	t	c	0	0	1	0	1
B4-1	c	g	a	g	g	act	c	c	c	a	c	c	c	t	t	c	t	a	g	c	c	1	0	0	0	1
C1-1	c	g	-	c	g	act	c	c	c	a	c	c	c	t	t	c	t	a	g	c	c	11	0	4	12	27
C2-1	c	g	-	c	g	act	c	c	c	a	c	c	c	c	t	c	t	a	g	c	c	10	0	2	0	12
C3-1	c	g	-	c	g	act	c	t	c	a	c	c	c	t	t	c	t	a	g	c	c	8	21	3	1	33
C4-1	c	g	-	c	g	act	c	t	c	a	t	c	c	t	t	c	t	a	g	c	c	1	0	0	0	1
Overall																						50	48	52	72	222

Note: Reference positions according to Yamamoto *et al.* (1984), positions in boldface are located within *Alu* U. Afr, African; Asn, Asian; Cauc, Caucasian, Ame, Amerind.

Table 2 Genetic diversity for each continent and for the worldwide sample

	n	Nhap[1]	H (SE) [2]	π (SE)[2,3]	θ_w/site[3]	Tajima's D	Fu & Li's D^*	Fu & Li's F^*
African	50	10	0.860 (0.020)	0.422 (0.046)	0.373	0.40	0.49075	0.54341
Asian	48	8	0.751 (0.048)	0.545 (0.029)	0.376	1.36	−0.00424	0.53249
Caucasian	52	9	0.716 (0.043)	0.503 (0.027)	0.369	1.07	−0.0344	0.40174
Amerind	72	6	0.755 (0.027)	0.492 (0.011)	0.265	2.30[4]	0.72411	1.49339[5]
World	222	16	0.831 (0.012)	0.522 (0.010)	0.409	0.73	−1.5052	−0.74865

Note: [1]Nhap, number of haplotypes; [2]SE, standard error, [3]values × 100; [4]$p < 0.05$; [5]$0.05 < p < 0.10$.

the *LDLR* gene is thus among the most variable nuclear (autosomal or *X*-linked) human loci investigated thus far. As expected from its presumed high evolutionary rate, the *Alu* U insertion contains most of the nucleotide variability of the region ($\pi = 0.973\% \pm 0{,}021\%$ *vs.* $\pi = 0.262\% \pm 0.009\%$ for the non-*Alu* U sites).

Considering all sites, there is no difference in nucleotide diversity between continents at the 95% confidence level. In contrast to many previous molecular genetic studies (Przeworski *et al.* 2000; Tishkoff & Verrelli, 2003) π reaches its maximum value in Asia, rather than Africa (whose mean diversity is only higher than America). But when H is used for comparison, Africa shows the highest value, which is statistically different from those of Europe and America, but not from Asia.

The AMOVA analysis using the Kimura-2P distance showed that only 7.9% of the genetic variation is found among continents, while the remaining 92.1% is found within continents. Grouping the data into Africans and Non-Africans results in a lower value for the among groups component (5.9%). Although all populations can be differentiated by their haplotype frequencies ($p < 0.05$ for all comparisons), the level of genetic variation between continents is low, and the overall F_{st} is 0.079, which is statistically different from zero, and is lower then the typical value of ∼0.15 observed for most nuclear genes (Tishkoff & Verrelli, 2003).

Considering a stationary population history, there is no deviation from neutral expectations for the whole sample, as measured from Tajima's D and Fu and Li's D^* and F^* (Table 2). When each continent is analyzed separately, however, there is a significantly positive D value ($D = 2.3$; $p < 0.05$) and a marginally non-significant F^* value ($F^* = 1.49$; $p < 0.10$) for America. These positive values might reflect the putative bottleneck associated with the initial peopling of the continent ∼20,000 years ago (Bonatto & Salzano, 1997), but could also be produced by balancing selection.

A completely different pattern, however, emerges when ancient population growth is taken into account. For both the whole sample and for each continent separately neutrality is strongly rejected for a wide range of scenarios (Fig. 2). The values obtained are higher than expected, suggesting either a population bottleneck or the action of balancing selection. Because all continental samples except America displayed the same pattern for the neutrality values, we think it is difficult to reconcile the high diversity exhibited by this locus with a population bottleneck affecting all human populations, since it has long been argued that Africa has always maintained a large population size (reviewed in Tishkoff & Verrelli, 2003), and that all continents (except America) have experienced population growth considering both uniparental (reviewed in Excoffier, 2002), and autosomal markers (for example, Zhivotovsky *et al.* 2003).

Although we have simulated a wide range of scenarios (see *Subjects and Methods*), rejection of the neutral model occurs even for small values of growth and very recent times for the onset of growth (Fig. 2). For example, considering the whole sample, all scenarios whose growth onset is 30,000 years or earlier yield statistically significant values, independent of the neutrality statistic used (Table 3). Considering each continent separately, the same pattern is obtained. For the Native American sample, all simulations produce statistically significant values, what is expected since for this sample even a stationary model rejects the neutral model. For the African, Asian and European samples, neutrality is rejected for almost all scenarios of growth occurring prior to 50,000 years before the present (ybp), even for small magnitudes of growth. Therefore, as exemplified in Table 3, while neutrality is rejected for scenarios in

Table 3 *P*-values of the neutrality tests statistics considering ancient population growth[1]

	Early growth scenario[2]			Late growth scenario[3]		
	Tajima's *D*	Fu & Li's D^*	Fu & Li's F^*	Tajima's *D*	Fu & Li's D^*	Fu & Li's F^*
African	0.008	0.006	0.006	0.087	0.063	0.063
Asian	0.005	0.003	0.003	0.034	0.096	0.096
Caucasian	0.001	0.002	0.002	0.045	0.102	0.102
Amerind	<0.001	<0.001	<0.001	0.001	<0.001	<0.001
World	<0.001	<0.001	<0.001	0.002	0.006	0.007

Note: [1]Values lower than 0.050 were considered statistically significant [2]Assuming a population growth starting 100,000 years ago and an expansion by a factor of 100. [3]Assuming a population growth by a factor of 100 starting at 36,000 years ago for whole sample and for Africans, at 26,000 years ago for Caucasians, at 18,000 years ago for Asians (Zhivotovsky et al., 2003), and at 20,000 years ago for Amerinds (S. L. B. unpublished data).

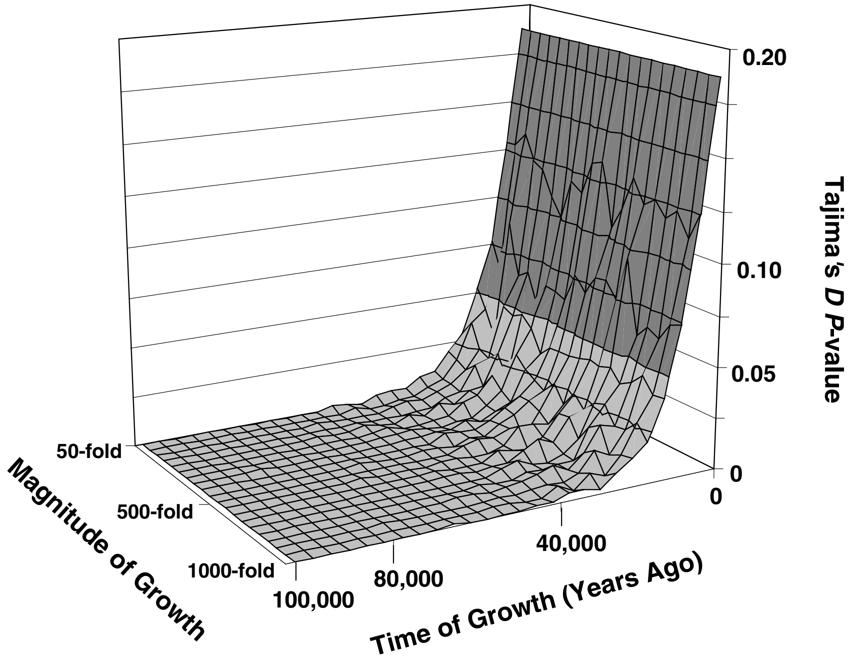

Figure 2 *P*-values of Tajima's *D* statistic under varying population history parameters. Light grey areas (*P*-values below 0.05) indicate a combinations of parameters under which the hypothesis of neutrality is rejected. Values obtained for ages of growth between 100,000 and 200,000 years ago (highly significant - *P*-values below 0.001) were omitted for clarity purposes. The same pattern was obtained when considering Fu & Li's D^* and F^* statistics, as well as each population separately (data not shown).

which growth started ~100,000 ybp (Wooding *et al.* 2004 and references therein), it is not for scenarios that assume a late growth for each continent (as in Zhivotovsky *et al.* 2003).

Haplotype Network

The evolutionary network of the inferred haplotypes is shown in Fig. 3. The position of the root was further confirmed by Maximum Parsimony analyses using the MEGA 2 program. Three clusters, A, B, and C, are observed, clusters B and C being very closely related, while A is the most differentiated, separated by at least six mutations from haplotype B4-1, its closest neighbour.

Confirming the results obtained for the AMOVA and F_{st} analyses, there is no obvious geographical

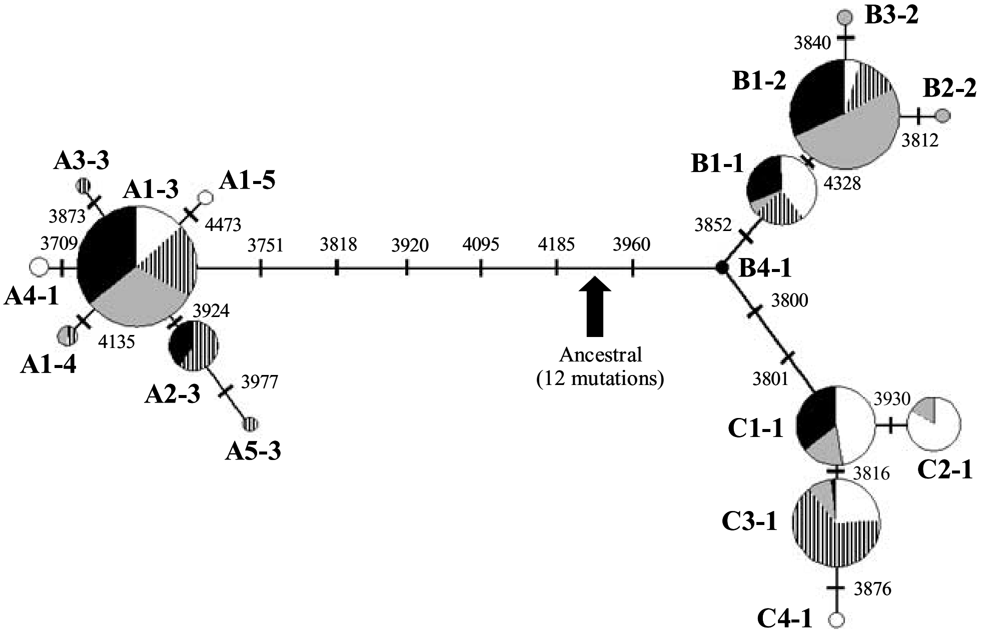

Figure 3 Median-joining network of the inferred haplotypes. The size of the circle is proportional to the haplotypes frequency in the whole sample. White represents African haplotypes, gray represents Asian, the hatched pattern represents European, and black represents Native American. The position of the outgroup, distant by 12 mutations, is indicated by the arrow.

structuring of haplotypes, and there is extensive sharing of haplotypes among different continents, particularly for the more common ones. The Amerind sample contains only the more common "node" haplotypes, in accordance with the lower level of haplotype diversity exhibited by this sample.

Coalescent Analysis

The T_{MRCA} and effective population size (N_e) of the sample was obtained from *Alu* U and non-*Alu* U sites independently, because of their different evolutionary rates. The estimates of θ obtained with Genetree were equal to θw for both sets of data, which yield similar estimates for both N_e and T_{MRCA}. *Alu* U sites resulted in a N_e estimate of 13,003 individuals, and in a T_{MRCA} estimate of 1,230,000 ybp, while non-*Alu* U sites furnished a N_e estimate of 10,194 individuals and a T_{MRCA} estimate of 1,080,000 years. Average values for N_e and T_{MRCA} were 11,598 individuals and 1,155,000 ybp, respectively. In this respect, our data agree with those of several other nuclear loci, which estimate the human N_e to be close to 10^4 (Chen & Li, 2001; Takahata *et al.* 2001; Tishkoff & Verrelli, 2003). The T_{MRCA} estimate is also in good agreement with most nuclear data (reviewed in Excoffier, 2002; Tishkoff & Verrelli, 2003).

Some of our analyses suggest balancing selection acting on this locus. Given the high frequency in all continents of Cluster A, the understanding of its origin may be of relevance to better comprehend the putative selective forces acting at *LDLR*. We have therefore used coalescent simulations to estimate the T_{MRCA} for Cluster A. Analysis of *Alu* U and non-*Alu* U sites gave a result of 616,000 and 462,000 ybp, respectively, with an average of 539,000 ybp. This value suggests a T_{MRCA} before the origin of anatomically modern humans. However, the coalescent models developed to date are still limited for the analysis of neutral genetic variation, which may not be the case for *LDLR*. If balancing selection is indeed affecting *LDLR* evolution, it would extend the coalescence times and it would therefore be more appropriate to take our T_{MRCA} estimates as the lower bound of the correct values.

Discussion

Evolutionary Rate and Nucleotide Diversity Estimates

The estimated non-*Alu* U sites substitution rate of 0.166% per Myr is in the expected range for an *Alu*-rich region, being exactly the same as previously reported for this family of elements (Chen & Li, 2001). However, the *Alu* U estimated rate of 0.632% per Myr is to our knowledge the highest estimated for an autosomal locus in humans (for a list of studies on this topic see Tishkoff & Verrelli, 2003). It is therefore important to critically evaluate the accuracy of this estimate. We are confident that our calculated rate is accurate for three reasons: first, the *Alu* Yb8 master sequence (the ancestral sequence) is well defined (Batzer *et al.* 1996); second, the evolutionary rate for CONSBI sites, as well as the conserved positions within the *Alu* insert, have been well studied (Britten, 1994); and third, the age estimated for the insertion of the *Alu* element in the *LDLR* locus agrees well with age estimates obtained in a comprehensive Alu Yb8 subfamily analysis (Carroll *et al.* 2001). Moreover, even if an extremely conservative calibration time of six million years for *Alu* U insertion is used, its evolutionary rate would be 0.347% per Myr, still well above the average *Alu* genomic rate of 0.166% per Myr (Chen & Li, 2001).

With respect to the nucleotide diversity, the same picture emerges; non-*Alu* U sites evolve at the average genomic rate, while *Alu* U harbours an unusual amount of variation. *Alu*-rich regions are known to evolve 1.61x faster then *Alu*-poor loci (Chen & Li, 2001). However, we think it unlikely that the higher *Alu* U diversity is caused only because of this. The *Alu* U evolutionary rate is 3.81x higher than the *Alu* average, and its nucleotide diversity is 3.71x higher than that obtained for non-*Alu* U sites (which contain *Alu* D and the partial *Alu*). Thus, *Alu* U seems to be an outlier even among other *Alu*-rich regions.

Genetic Evidence for Balancing Selection?

Under balancing selection, high genetic diversity is expected both at the target sites as well as at linked sites. A puzzling pattern displayed by the studied region is that although *Alu* U is highly diverse, non-*Alu* U sites have diversity values similar to other autosomal loci studied so far. A possible explanation for this is that, given the lower evolutionary rate of non-Alu U sites (compared to Alu U), the time elapsed since the emergence of balancing selection was not enough to result in a higher diversity. Interestingly, Baum *et al.* (2002) found a similar pattern in *GYPA,* in which one exon displayed an eight-fold higher diversity than that of its surrounding region ($\pi = 2.4\%$ for exon 2 *vs.* $\pi = 0.3\%$ for the entire region).

It is worthwhile considering if the higher *Alu* U diversity could have been shaped by gene conversion, as this mechanism has been suggested to have played a central role in *Alu* evolution (Batzer & Deininger, 2002). However, we think gene conversion is not a likely explanation for our data for the following reasons. If the gene conversion has involved *Alu* elements from different subfamilies, we would expect that some of the *LDLR* haplotypes would lack Yb8 diagnostic positions, having diagnostic sites for other subfamilies, but this is not the case for our data. Additionally, we would expect both a deeper divergence time for *Alu* U than for the non-*Alu* U sites, as well as the occurrence of tightly linked sets of SNPs in a given clade. However, the T_{MRCA} for both *Alu* U and non-*Alu* U sites are in very close agreement, and *Alu* U variation is dispersed along the whole element.

Besides the high diversity shown by this locus, the low overall F_{st}, the low between-group component of the AMOVA when Africans and non-Africans are compared, the absence of any haplotype substructure in the haplotype network, and the same levels of nucleotide diversity in both African and non-African populations, are suggestive patterns of balancing selection, especially if the selective pressures are the same for both African as well as non-African populations.

Although none of the neutrality tests were able to reject it when a stationary population is considered, they are all significant for a wide range of growth scenarios, being non-significant only when the onset of growth is recent (50,000 ypb or less), and the magnitude of growth is low. Ignoring past human population history will cause ordinary neutrality tests to be too conservative in detecting balancing selection, and too liberal in detecting positive selection, since the ancient population expansion will drive the neutrality estimators at all loci to more negative values. While it seems clear that these neutrality tests are highly sensitive to the prior demographic history of the studied populations for detecting natural selection, and that this should be accounted for (Kreitman & Di Rienzo, 2004; Wooding *et al.*, 2004), a much more troubling issue is what demographic scenario is to be assumed in order to correct the significance of these statistics. In Table 3, we compare the *P*-values obtained considering early (Wooding *et al.* 2004), or late (Zhivotovsky *et al.* 2003) population growth. For the whole sample, and for the American sample, neutrality is rejected in both scenarios. However, for the Asian and Caucasian samples, neutrality was not rejected for the late growth scenario considering Fu & Li's D^* and F^*, and for the African sample neutrality was not rejected for any statistics considering the late growth scenario (Table 3).

Recently, Zhang *et al.* (2003) reported the high abundance in the genome of several organisms, including humans, of highly divergent, frequent haplotypes, which were called yin yang haplotypes. In their study, they show that about 80% of the yin yang haplotypes could have arisen by pure neutral mechanisms. Adopting Zhang *et al.* (2003) criteria, the whole cluster A, on one hand, and haplotypes C3-1 and C4-1, on the other, consist of a yin yang pair. These results are of relevance to show that highly divergent and frequent haplotypes might be present in the population only by neutral evolution. However, the suggestion for balancing selection in the 3′-UTR of the *LDLR* gene was based not on the presence of yin yang haplotypes, but essentially on the outcome of several statistical methods that seem to point in this direction.

Since the present results indicate balancing selection as a possible evolutionary force over this locus, it should also be considered whether the 3′-UTR of the *LDLR* gene could regulate gene expression. There is accumulating data suggesting that *LDLR* gene expression can be regulated by post-transcriptional mechanisms. Phorbol esters (Wilson *et al.* 1997), Genfibrozil (Goto *et al.* 1997) and chenodeoxycholic acid (Nakahara *et al.* 2002) seem to modulate the stability of *LDLR* mRNA, enhancing its half-life. Although the molecular mechanism for this effect is still unresolved, the proposed model involves the association of mRNA with cytoskeletal elements through its 3′-UTR, with the *Alu*-rich region being proposed as a cytoskeleton binding domain (Wilson *et al.* 1998). A direct involvement of the *LDLR* 3′-UTR on the gene expression was further demonstrated by Knouff *et al.* (2001), who showed that deletion of this region doubled the expression of *Ldlr* in mice.

Acknowledgments

We thank Roberta Eckert for expert technical assistance, Lynn B. Jorde and David Witherspoon for providing some DNA samples, and L. B. J., D. W., Felipe G. Grazziotin, and Eduardo Eizirik for critical reading, as well as two anonymous reviewers for several suggestions on an early version of the manuscript, and C. E. A. Coimbra, Ricardo V. Santos, Kim Hill, and Madaglena Hurtado for assistance in the Amerindian field collections. The latter have been approved by the Brazilian Ethics National Committee (CONEP, resolution 123/98). This work was supported by grants from Programa de Apoio a Núcleos de Excelência (PRONEX) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

References

Bamshad, M. & Wooding, S. P. (2003) Signatures of natural selection in the human genome. *Nat Rev Genetics* **4**, 99–111.

Bamshad, M. J., Mummindi, S., Gonzales, E., Ahuja, S. S., Dunn, D. M., Scott Watkins, W., Wooding, S., Stone, A. C., Jorde, L. B., Weiss, R. B. & Ahuja, S. K. (2002) A strong signature of balancing selection in the 5′ *cis*-regulatory region of *CCR5*. *Proc Natl Acad Sci USA* **99**, 10539–10544.

Bandelt, H-J., Forster, P. & Röhl, A. (1999) Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* **16**, 37–48.

Batzer, M. A. & Deininger, P. L. (2002) *Alu* repeats and human genomic diversity. *Nat Rev Genetics* **3**, 370–380.

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

Baum, J., Ward, R. H. & Conway, D. J. (2002) Natural selection on the erythrocyte surface. *Mol Biol Evol* **19**, 223–229.

Bonatto, S. L. & Salzano, F. M. (1997) Diversity and age of the four major mtDNA haplogroups, and their implications for the peopling of the New World. *Am J Hum Genet* **61**, 1413–1423.

Britten, R. J. (1994) Evidence that most human *Alu* sequences were inserted in a process that ceased about 30 million years ago. *Proc Natl Acad Sci USA* **91**, 6148–6150.

Carroll, M. L., Roy-Engel, A. M., Nguyen, S. V., Salem, A-H., Vogel, E., Vincent, B., Myers, J., Ahmad, Z., Nguyen, L., Sammarco, M., Watkins, W. S., Henke, J., Makalowski, W., Jorde, L. B., Deininger, P. L., Batzer, M. A. (2001) Large-scale analysis of the *Alu* Ya5 and Yb8 subfamilies and their contribution to human genomic diversity. *J Mol Biol* **311**, 17–40.

Chen, F-C. & Li, W-H. (2001) Genomic divergences between humans and other hominoids and the effective population size of common ancestors of humans and chimpanzees. *Am J Hum Genet* **68**, 444–456.

Clark, A. G., Weiss, K. M., Nickerson, D. A., Taylor, S. L. & Buchanan, A. (1998) Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am J Hum Genet* **63**, 595–612.

Excoffier, L. (2002) Human demographic history: refining the recent African origin model. *Curr Opin Genet Dev* **12**, 675–682.

Excoffier, L., Smouse, P. & Quattro, J. (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131**, 479–491.

Fu, Y-X. (1997) Statistical tests of neutrality of mutations against population growth, genetic hitchhiking and background selection. *Genetics* **147**, 915–925.

Fu, Y-X. & Li, W-H. (1993) Statistical tests of neutrality of mutations. *Genetics* **133**, 693–709.

Goldstein, J., Hobbs, H. & Brown, M. (1995) Familial hypercholesterolemia. In: *The Metabolic Basis of Inherited Disorders* (eds.C. R. Scriver, A. L. Beaudet, W. S. Sly & D. Valle), pp. 1981–2038, McGraw-Hill, New York.

Goto, D., Okimoto, T., Ono, M., Shimotsu, H., Abe, K., Tsujita, Y. & Kuwano, M. (1997) Upregulation of low density lipoprotein receptor by genfibrozil, a hypolipidemic agent, in human hepatoma cells through stabilization of mRNA transcripts. *Arterioscl Thromb Vasc Biol* **17**, 2707–2712.

Griffiths, R. C. & Tavare, S. (1994) Sampling theory for neutral alleles in a varying environment. *Proc R Soc London B Biol Sci* **344**, 403–410.

Harding, R. M., Fullerton, S. M., Griffiths, R. C., Bond, J., Cox, M. J., Schneider, J. A., Moulin, D. S. & Clegg, J. B. (1997) Archaic African *and* Asian lineages in the genetic ancestry of modern humans. *Am J Hum Genet* **60**, 772–789.

Harris, E. & Hey, J. (1999) X chromosome evidence for ancient demographic human histories. *Proc Natl Acad Sci USA* **96**, 3320–3324.

Kass, D. H., Batzer, M. A. & Deininger, P. L. (1995) Gene conversion as a secondary mechanism of short interspersed element (SINE) evolution. *Mol Cell Biol* **15**, 19–25.

Knouff, C., Malloy, S., Wilder, J., Altenburg, M. K. & Maeda, N. (2001) Doubling expression of the low density lipoprotein receptor by truncation of the 3′-untranslated region sequence ameliorates type III hyperlipoproteinemia in mice expressing the human ApoE2 isoform. *J Biol Chem* **276**, 3856–3862.

Koda, Y., Tachida, H., Soejima, M., Takenaka, O. & Kimura, H. (2004) Population differences in DNA sequence variation and linkage disequilibrium at the *PON1* gene. *Ann Hum Genet* **68**, 110–119.

Kumar, S., Tamura, K., Jakobsen, I. B. & Nei, M. (2001) MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics* **17**, 1244–1245.

Kreitman, M. & Di Rienzo, A. (2004) Balancing claims for balancing selection. *Trends Genet* **20**, 300–304.

Labuda, D., Zietkiewicz, E. & Mitchell, G. A. (1995) *Alu* elements as a source of genomic variation: deleterious effects and evolutionary novelties. In: *The Impact of Short Interspersed Elements (SINEs) on the Host Genome* (ed R. J. Maraia), pp. 1–24, Springer, R.G. Landes, New York.

Lehrman, M. A., Russell, D. W., Goldstein, J. L. & Brown, M. S. (1987) *Alu-Alu* recombination deletes splice acceptor sites and produces secreted low density lipoprotein receptor in a subject with familial hypercholesterolemia. *J Biol Chem* **262**, 3354–3361.

Nachman, M. W. & Crowell, S. L. (2000) Contrasting evolutionary histories of two introns of the Duchenne muscular dystrophy gene, *Dmd*, in humans. *Genetics* **155**, 1855–1864.

Nakahara, M., Fujii, H., Maloney, P. R., Shimizu, M. & Sato, R. (2002) Bile acids enhance low density lipoprotein gene expression via a MARK cascade-mediated stabilization of mRNA. *J Biol Chem* **277**, 37229–37234.

Nei, M. & Kumar, S. (2000) *Molecular Phylogenetics and Evolution*, Oxford Univ Press, New York.

Posada, D. & Crandall, K. A. (1998) Modeltest: testing the model of DNA substitution. *Bioinformatics* **14**, 817–818.

Przeworski, M., Hudson, R. R. & Di Rienzo, A. (2000) Adjusting the focus on human variation. *Trends Genet* **16**, 296–302.

Raymond, M. & Rousset, F. (1995) An exact test for population differentiation. *Evolution* **49**, 1280–1283.

Rozas, J. & Rozas, R. (1999) DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* **15**, 174–175.

Ruhlen, M. (1994) *The origin of language*, John Wiley and Sons, New York.

Sachidanandam, R., Weissmann, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., Mullikin, J. C., Mortimore, B. J., Willey, D. L., Hunt, S. E., Cole, C. G., Coggill, P. C., Rice, C. M., Ning, Z., Rogers, J., Bentley, D. R., Kwok, P. Y., Mardis, E. R., Yeh, R. T., Schultz, B., Cook, L., Davenport, R., Dante, M., Fulton, L., Hillier, L., Waterston, R. H., McPherson, J. D., Gilman, B., Schaffner, S., Van Etten, W. J., Reich, D., Higgins, J., Daly, M. J., Blumenstiel, B., Baldwin, J., Stange-Thomann, N., Zody, M. C., Linton, L., Lander, E. S. & Altshuler, D. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933.

Schneider, S., Roessli, D. & Excoffier, L. (2000) Arlequin ver.2.000: a software for population genetics data analysis. Genetics and Biometry Lab, Department of Anthropology, University of Geneva, Switzerland.

Stephens, M., Smith, M. J. & Donnelly, P. (2001a) A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* **68**, 978–989.

Stephens, J. C., Schneider, J. A., Tanguay, D. A., Choi, J., Acharya, T., Stanley, S. E., Jiang, R., Messer, C. J., Chew, A., Han, J. H., Duan, J., Carr, J. L., Lee, M. S., Koshy, B., Kumar, A. M., Zhang, G., Newell, W. R., Windemuth, A., Xu, C., Kalbfleisch, T. S., Shaner, S. L., Arnold, K., Schulz, V., Drysdale, C. M., Nandabalan, K., Judson, R. S., Ruano, G. & Vovis, G. F. (2001b) Haplotype variation and linkage disequilibrium in 313 human genes. *Science* **293**, 489–493.

Stiner, M. C., Munro, N. D., Surovell, T. A., Tchernov, E. & Bar-Yosef, O. (1999) Paleolithic population growth pulses evidenced by small animal exploitation. *Science* **283**, 190–194.

Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595.

Takahata, N., Lee, S-H. & Satta, Y. (2001) Testing multiregionality of modern human origins. *Mol Biol Evol* **18**, 172–183.

Tishkoff, S. A. & Verrelli, B. C. (2003) Patterns of human genetic diversity: implications for human evolutionary history and disease. *Ann Rev Genom Hum Genet* **4**, 293–340.

Watterson, G. (1975) On the number of segregating sites in genetical models without recombination. *Theor Popul Biol* **7**, 256–276.

Wilson, G. M., Roberts, E. A. & Deeley, R. G. (1997) Modulation of LDL receptor cell stability by phorbol esters in human liver cell culture models. *J Lipid Res* **38**, 437–446.

Wilson, G. M., Vasa, M. Z. & Deeley, R. G. (1998) Stabilization and cytoskeletal-association of LDL receptor mRNA are mediated by distinct domains in its 3′untranslated region. *J Lipid Res* **39**, 1025–1032.

Wooding, S., Kim, U-K., Bamshad, M. J., Larsen, J., Jorde, L. B. & Drayna, D. (2004) Natural selection and molecular evolution in *PTC*, a bitter-taste receptor gene. *Am J Hum Genet* **74**, 637–646.

Wright, S. (1969) *Evolution and the genetics of populations: the theory of gene frequencies*. Vol. 2. *The theory of gene frequencies*, University of Chicago Press, Chicago.

Yamamoto, T., Davis, C. G., Brown, M. S., Schneider, W. J., Casey, M. L., Goldstein, J. L. & Russel, D. W. (1984) The human LDL receptor: a cysteine-rich protein with multiple *Alu* sequences in its mRNA. *Cell* **39**, 27–38.

Zhang, J., Rowe, W. L., Clark, A. G., Buetow, K. H. (2003) Genomewide distribution of high-frequency, completely mismatching SNP haplotype pairs observed to be common across human populations. *Am J Hum Genet* **73**, 1073–1081.

Zhivotovsky, L. A., Rosenberg, N. A. & Feldman, M. W. (2003) Features of evolution and expansion of Modern Humans, inferred from genomewide microsatellite markers. *Am J Hum Genet* **72**, 1171–1186.

Received: 2 July 2004
Accepted: 17 November 2004