

Rene J. Herrera · Robert K. Lowery · Abraham Alfonso
John F. McDonald · Javier R. Luis

Ancient retroviral insertions among human populations

Received: 17 November 2005 / Accepted: 25 December 2005 / Published online: 4 March 2006
© The Japan Society of Human Genetics and Springer-Verlag 2006

Abstract Human endogenous retroviruses (HERVs) represent vestiges of ancient infections that resulted in stable integration of the viral genome. These insertional elements of viral origin are in fact molecular fossils and, as such, a source of evolutionary information. A family of HERV insertions designated HERV-K includes members that are still polymorphic for the original insertional event. The goal of this report is to describe a novel genetic marker system based on polymorphic retroviral insertions (PRVIs) and to assess its potential usefulness in human population genetic analyses. The allelic frequencies of four insertionally polymorphic HERV-K loci were analyzed in nine geographically targeted, worldwide populations. A polymerase chain reaction assay was employed to examine the frequencies of the provirus and/or solo long terminal repeat insertions at these four loci. Several statistical and phylogenetic analyses were performed based on the frequency data. The phylogenetic relationships observed among the nine populations based on the four retroviral HERV-K loci are consistent not only with prior genetic analyses with other traditional marker systems but also with reported historical and biogeographical data. These polymorphic endogenous retroviral sequences display features that make them excellent tools for forensic and population genetic studies.

Keywords HERV-K · PRVI · HML-2 · Endogenous retrovirus · Human evolution · Forensic · Genetic diversity

Introduction

Human endogenous retroviruses (HERVs) are transposable genetic elements that are believed to be the remnants of exogenous retroviral infections that, subsequent to insertion, amplify and disperse (Von Sternberg et al. 1992; Leib-Mosch and Seifarth 1996; Löwer et al. 1996; Patience et al. 1997). Sequence analysis of different retroelements suggests that approximately 40 million years ago, prior to the separation of Old World and New World monkeys, the primate germline experienced a sharp peak in retroviral integrations, yet currently the rate of insertion is much lower (Lander et al. 2001). At least 22 distinct HERV families have been identified (Tristem 2000) occupying approximately 7% of the human genome (Smit 1996; Bock and Stoye 2000). Their prototypical structure contains *gag*, *pol*, and *env* genes flanked by two identical long terminal repeats (LTRs). However, HERV elements are often truncated, and in many cases only one LTR (solo-LTR) remains as vestige of the original provirus. Recombination involving the LTRs usually results in the deletion of the three viral open reading frames and one of the two LTRs. The persisting LTR is the solo-LTR.

It has been suggested that HERVs may contribute to the development of cancer and autoimmune diseases in humans (Löwer 1999; Dunn et al 2003). Nonpathogenic effects of endogenous retroviruses, such as protection against exogenous retroviral infection (Löwer 1999; Sverdlov 2000; Dunn et al 2003) and participation in the formation of the placenta (Villareal 1997; Chen et al. 2000), have also been proposed. In addition, transposable elements, like HERVs, have been shown to lead to increased genetic variability. Thus, it is possible that they may play a role in the evolution of the host

R.J. Herrera (✉) · R.K. Lowery · A. Alfonso · J.R. Luis
Department of Biological Sciences,
Florida International University, University Park,
OE 304, Miami, FL 33199, USA
E-mail: herrerar@fiu.edu
Tel.: +1-305-3481258
Fax: +1-305-3481259

J.R. Luis
Departamento de Bioquímica, Xenética e Inmunoloxía,
Universidade de Vigo, Vigo, Spain

J.F. McDonald
School of Biology, Georgia Institute of Technology,
310 Ferst Drive, Atlanta, GA 30332-0230, USA

organism (McDonald 1995; Brosius 1999). LTRs contain control elements for the expression of DNA and cases have been described where HERV elements affect the expression of adjacent genes (Brosius 1999). The fortuitous insertion of these elements throughout the genome may also have had a significant impact on the evolution of the human lineage (Hughes and Coffin 2004).

The LTRs at either end of a provirus are identical at the time of insertion. Therefore, the degree of sequence divergence between any such pair of LTRs can be used as a measure of the time elapsed since the insertion event. Various methods of dating HERV insertions have been developed to reconstruct interspecific phylogeny (Medstrand and Mager 1998; Barbulescu et al. 1999; Johnson and Coffin 1999; Hughes and Coffin 2004). One method of dating (Johnson and Coffin 1999) estimates the time of integration on the basis of the sequence differences between 5' and 3' LTRs. The rate of substitution employed in these calculations are in the order of 2.3×10^{-9} to 5×10^{-9} substitutions per site per year (Johnson and Coffin 1999). One of the subdivisions within the HERV viruses is the HERV-K family. While the HERV-K family is estimated to be as much as 50 million years old (Lavie et al. 2004), the HML-2 subfamily of HERV-K dates to roughly 28 million years ago (Reus et al. 2001). Members of the HML-2 subfamily are seen in Old World monkeys, apes and humans, but not in New world monkeys. Phylogenetic trees based on sequence data from HERV-K elements of primate species generate topologies consistent with established evolutionary phylogeny (Johnson and Coffin 1999; Jordan and McDonald 2002). Several studies indicate that the spread and amplification of distinct HERV-K families and subfamilies have occurred at different times during the evolution of the lineage leading to modern humans (Wilkinson et al. 1994; Medstrand and Mager 1998; Sverdlov 2000; Jordan and McDonald 2002).

The HML-2 subfamily has been phylogenetically examined (Medstrand and Mager 1998; Turner et al. 2001; Budzin et al. 2003; Macfarlane and Simmonds 2004) resulting in the subdivision of this HML-2 group into several subtypes. One of these subtypes of HML-2, Cluster 9 in Medstrand and Mager (1998), had its origin about the time of separation of the human and the Great Ape lineages, and seems to encompass the most recent of the HERV-K integrations. All but one of the known members of this subtype are human-specific. Some human-specific HML-2 elements have been reported to exhibit insertional polymorphisms for both full-length proviral elements and/or solo-LTRs (Hughes and Coffin 2004). Recently, polymorphic HERV-K loci have been employed in a number of human population studies. For example, phylogenetic analyses using HERV-K polymorphisms have corroborated the greater genetic diversity in sub-Saharan Africa (Macfarlane and Simmonds 2004). Mamedov et al. (2004) have also used these markers to characterize Russian populations and

to contrast them to African groups. Autosomal haplogroups involving single nucleotide polymorphisms (SNPs) within HERV-K open reading frames (ORFs) have also been used to study several human populations (Mayer et al. 2005).

Insertionally polymorphic HERV elements possess three attributes of value for human phylogenetic analyses. First, the age of insertions can be estimated by assessing the nucleotide sequence divergence of the homologous LTRs. Second, the ancestral state, the lack of insertion, is known. Third, since HERV integrations are unique and irreversible events, individuals sharing any given insertion are related by descent not just by state. Even if a provirus is truncated, it usually leaves behind sequence vestiges in the form of solo-LTRs.

The goal of this study is to ascertain the potential usefulness of polymorphic HERV-K insertions in human population genetic analyses. Nine geographically targeted, worldwide human populations were examined for their allelic frequencies at four polymorphic retroviral insertions (PRVIs). The elements studied were HERV-K 113 (Turner et al. 2001), HERV-K 115 (Turner et al. 2001), Z80898a (Medstrand and Mager 1998) and HERV-K 11q22 (Hughes and Coffin 2004). The HERV-K 113 and HERV-K 115 insertions represent full-length proviruses while Z80898a is a solo-LTR remnant. The 11q22 PRVI exhibits two variants of the proviral insertion as well as three versions of the solo-LTRs. In the present study, we only scored the 11q22 insertions as provirus and solo-LTR. By analyzing nine worldwide populations for these polymorphic HERV-K insertions, we are introducing a new system for studying human phylogenies.

Materials and methods

DNA samples and populations

Samples of blood were taken from healthy, unrelated, autochthonous individuals from the populations of the Basque Country in Spain, Cameroon, China, Egypt, Galicia in Spain, Kenya, Oman, Rwanda, and Taiwan (Ami aborigines). The samples were collected as whole blood in EDTA vacutainer tubes in adherence with the guidelines set forth by Florida International University's Internal Review Board. Table 1 illustrates the geographical and linguistic classification of the populations studied. DNA was extracted using the phenol-chloroform method (Sambrook et al. 1989) and stored at -70°C .

DNA amplification

Using the polymerase chain reaction (PCR), each DNA sample was amplified for the HERV-K insertional loci K113, K115, 11q22 and Z80898a (see Table 2 for accession numbers). With the K113 and K115 PRVIs, a

Table 1 Description of the populations studied and insertion frequencies at the four loci analyzed (*11Q22*, *K113*, *K115*, *Z80898a*). *ProV* Provirus, *Solo* solo LTR

Population	Description	Linguistic Classification ^a		N	11Q22		N	K113	N	K115	N	Z80898a
		Level	Sublevel		ProV	Solo						
sub-Saharan Africa												
Kenya	Bantu	Niger-Congo	Benue-Congo	46	0.576	0.261	46	0.359	50	0.22	44	0.307
Rwanda	Hutu/Tutsi	Niger-Congo	Benue-Congo	65	0.577	0.308	49	0.224	49	0.224	43	0.302
Cameroon	Bamileke/Bantu	Niger-Congo	Benue-Congo	15	0.467	0.233	16	0.094	16	0.219	16	0.094
Middle-East												
Egypt	Arabs/Berbers	Afro-Asiatic	Semitic	52	0.615	0.346	43	0.058	45	0.133	44	0.432
Oman	Arabs	Afro-Asiatic	Semitic	52	0.558	0.394	43	0.035	57	0.149	52	0.327
Southwest Europe												
Galicia	Galician/Celtic	Indo-European	Italic	42	0.381	0.619	48	0.021	50	0.07	45	0.422
Basque	Basque	Basque	Basque	44	0.727	0.273	50	0.02	49	0.051	47	0.479
East Asia												
China	Han Chinese	Sino-Tibetan	Chinese	39	0.077	0.82	44	0.125	42	0.036	31	0.516
Taiwan, Ami	Ami	Austronesian	Formosan	54	0	0.926	47	0.16	49	0.041	45	0.656

^aInformation on linguistic classifications was obtained from <http://www.ethnologue.com>

Table 2 Primer sequences used for DNA amplifications. *LTR* Long terminal repeat

Primer	Sequence	Reference	Accession number
11Q22 5'	5'-TCTAGTGTATCTGATTCTCACC-3'	Hughes and Coffin 2004	AP000776
11Q22 3'	5'-TCCACTTCATTAACCTGTCAAC-3'	Hughes and Coffin 2004	AP000776
K115 5'	5'-CCGCACCTAGTCAACTTAGC-3'	Turner et al. 2001	AF189745
K115 3'	5'-CCGCACCTAGTCAACTTAGC-3'	Turner et al. 2001	AF202031
K113 5'	5'-GCATGGGGAGATTCAGAACC-3'	Turner et al. 2001	AC068624
K113 3'	5'-CATGTTTCCTGTCGCCCCAC-3'	Turner et al. 2001	AC068624
Z80898a 5'	5'-TAGAAGTAGCTCTGTTCTGTGG-3'	Medstrand and Mager 1998	Z80898
Z80898a 3'	5'-CTACCAGTCCAAACCATTTCAGT-3'	Medstrand and Mager 1998	Z80898
Z80898a 3'II	5'-TGGAGTAGCACAATCATTACC-3'	Present study	Z80898
LTR	5'-CCTCCTCAGCACAGACCCTTTAC-3'	Medstrand and Mager 1998	AC002400
Control	5'-CCCAATATTAAGATCACATCCTTAGC-3'	Present study	AC002400

single amplification reaction for each locus was performed using a primer unique for the sequence flanking the insertion and a consensus LTR primer based on a sequence present in all LTRs studied here. For 11q22, two reactions were needed. In the first PCR, flanking primers were used to detect the lack of insertion and/or the presence of solo-LTR alleles. A second amplification reaction was then conducted using the 3' flanking primer and the consensus LTR primer to detect the presence of the provirus insertion. For the Z80898a locus, an initial PCR was also done using flanking primers to amplify the lack of insertion allele and/or the solo-LTR as with 11q22. A second amplification reaction was then performed to avoid missing the insertion allele (solo-LTR) in heterozygous individuals due to the preferential amplification of the smaller allele (lack of insertion). In this confirmatory reaction, the presence or absence of the insertion was assayed using the consensus LTR primer and a 3' flanking primer (Z80898a) different from the one employed in the first reaction. An internal standard, consisting of primers corresponding to the 3' flanking region of a monomorphic HERV-K insertion and the consensus LTR was included in this amplification to detect false negatives. The sequences of primers

utilized in PCRs are given in Table 2. Amplifications were performed in a total volume of 15 µl containing 200 µM of each dNTP, 3 U *Taq* polymerase, 1.5 mM MgCl₂, 10 mM Tris-HCl (pH 8.3) and 50 mM KCl, in a Thermal Cycler 480 (Perkin Elmer, Wellesley, MA) with an initial denaturing step of 94°C for 4.5 min, then 35 cycles with the following parameters: 1 min at 95°C, 1.5 min at annealing temperature (see Table 3), 6 min at 72°C and a final extension step of 10 min at 72°C. Specific PCR conditions for the various amplifications are summarized in Table 3. The amplicons were electrophoresed in 2% agarose, 1X TAE gels with *HaeIII*-digested ϕ X174 DNA as size marker.

Data analysis

Allelic and genotypic frequencies were determined using the gene counting method (Li 1976). Observed and expected heterozygosities were calculated using the Biosys-2 program (Black 1997). The state of genetic equilibrium (Hardy-Weinberg) was evaluated by means of the Chi-square test with Bonferroni corrections. The degree of genetic differentiation among populations was assessed

Table 3 Polymerase chain reaction (PCR) conditions for the analysis of the four loci studied

	K115	K113	11q22 1st reaction	11q22 2nd reaction	Z80898a	Z80898a confirmation
Pre-denaturing temperature	94	94	95	95	94	94
Annealing temperature	64°C	64°C	60°C	55°C	62°C	58°C
Primers (ng/μl)	5': 7.5, 3': 2.5, LTR: 5	5': 7.5, 3': 2.5, LTR: 5	5	5	5	3' and control: 5, LTR: 10
MgCl ₂ (mM)	1.5	1.5	3.5	3.5	1.5	1.5

with the *G*-test (Carmody 1991). Power of discrimination (PD) was calculated following the method of Desmarais et al. (1998). Inter- and intra-population components of genetic variation were generated by means of the DISPAN program (Ota 1993). For this purpose, the populations were grouped according to geography as follows: Cameroon, Kenya, and Rwanda as sub-Saharan Africa (SS), Egypt and Oman as North Africa/Arabian Peninsula (NA/AP), Basque Country and Galicia as Europe (EU) and Taiwan Ami aborigines and China as East Asia (EA).

A maximum likelihood (ML) phylogenetic tree was derived from allelic frequencies using the PHYLIP package version 3.6 (Felsenstein 1989). A hypothetical outgroup homozygous for the lack of the insertion allele in all four loci was used for rooting the tree. Statistical support for the branch nodes of the dendrogram was ascertained using bootstrap analysis of 1,000 replicates. The relative degree of gene flow experienced and/or effective population size of the populations was estimated by plotting the distance from the centroid versus the heterozygosity as described by Harpending and Ward (1982). Principal component (PC) analysis was performed with the NTSYpc program version 2.02i (Rohlf 1993) to reveal the genetic relationship among the different populations.

To better understand the relationships between the four PRVIs examined in this study (K113, K115, 11q22 and Z80898a) and a group of previously reported fixed HERV-K insertions, a phylogenetic analysis of these two groups was conducted using a conserved 968 bp LTR sequence. The fixed HERV-K elements examined are human-specific and were previously reported in Medstrand and Mager (1998) as members of Cluster 9 in their dendrogram. For simplicity, we have designated all fixed solo LTRs by letter (fixed solo B–G). The GenBank accession numbers for these fixed loci are: fixed Solo B (AC002508), fixed Solo C (U73641a), fixed Solo D (L47334), fixed Solo E (Z84493), fixed Solo F (AC002350) and fixed Solo G (U47924). As per Medstrand and Mager (1998), a letter at the end of an accession number indicates the presence of more than one LTR within the entry (e.g., U73641a and Z80898a). To generate the dendrogram, sequences were aligned using Clustal-X, edited in McClade and used to construct a neighbor-joining (NJ) tree with absolute genetic distance. Bootstrapping was performed with 1,000 bootstrap replicates using PAUP, version 4.0. In order

to illustrate the distances between these sequences, NJ distances were utilized in Clustal-X to derive a PC plot using NTSYpc program version 2.02i.

Results

The insertional frequencies of the four HERV-K PRVIs examined in the nine populations are displayed in Table 1. The lowest frequency for the Z80898a insertion allele was found in Cameroon and the highest in the Ami from Taiwan with an overall occurrence of 39.3%. The insertion values at the Z80898a locus exhibits a trend from highest in East Asia, intermediate in Europe and the Middle East, to lowest in sub-Saharan Africa. On the other hand, the K-115 provirus is lowest in China (4%) and highest in Rwanda (22.4%) with an overall frequency in the nine populations of 12.7%. This clinal distribution is the inverse of that observed for Z80898a. In the case of K-113, the highest insertion frequency was found in Kenya (35.9%) and the lowest in Europe (~2.0%), with an overall frequency of 12.2%. Regionally, the K-113 insertion is highest in sub-Saharan East Africa (average 29.2%), intermediate in East Asia (average 14.3%), and lowest in the Middle East and Europe (average of 3.6%). The 11q22 proviral insertion was detected at lowest frequencies in East Asia (average 3.9%) and much higher outside of East Asia (average 55.7%). The highest frequency of the 11q22 solo-LTR allele was found in East Asia (average 87.3%), intermediate in Galicia (61.9) and lowest in sub-Saharan Africa (average 26.7%). The proportion of the 11q22 solo-LTR to full length provirus is lower than 1 in Africa, the Middle East and Europe (except in Galicia) and higher than 10 in East Asia.

The highest overall heterozygosity values are observed among the sub-Saharan group of populations; the Middle East, Europe, and East Asia follow in decreasing order (Table 4). All nine populations except Cameroon exhibit substantially lower than expected heterozygosities for the Z80898a locus. This observed heterozygosity deficit ranges from 75 to 33% of the expected values. The Chi-square test for Hardy–Weinberg equilibrium with Bonferroni corrections detected significant differences from the expected genotypic frequencies at Z80898a in the Basque Country, China, Egypt, Galicia, Rwanda and the Ami aborigines of Taiwan.

Overall PD values for the populations studied range between 0.3697 (K115) to 0.7323 (11q22) with an overall for all loci combined of 0.9591 (Table 5). The G-paired test (Table 6) displays evidence of genetic differentiation in all the pair-wise population comparisons except for

the Basque Country with Egypt ($p=0.079$), Egypt with Oman (0.685), and Kenya with Rwanda (0.255). In order to investigate the allocation of genetic variation at the intra- (*Hs*) and inter- (*Gst*) levels, the nine populations were analyzed as a whole set as well as subdivided

Table 4 Expected and observed heterozygosities for all four loci and all nine populations

	Kenya	Rwanda	Cameroon	Egypt	Oman	Basque	Galicia	China	Taiwan	Mean
11q22										
Expected	0.580	0.564	0.660	0.505	0.536	0.401	0.477	0.314	0.138	0.464
Observed	0.674	0.569	0.800	0.462	0.558	0.409	0.476	0.256	0.148	0.484
K113										
Expected	0.465	0.352	0.175	0.111	0.068	0.040	0.041	0.221	0.271	0.194
Observed	0.457	0.367	0.188	0.116	0.070	0.040	0.042	0.250	0.277	0.201
K115										
Expected	0.347	0.352	0.353	0.234	0.256	0.098	0.132	0.070	0.079	0.213
Observed	0.440	0.449	0.438	0.267	0.298	0.102	0.140	0.071	0.082	0.254
Z80898a										
Expected	0.430	0.427	0.175	0.496	0.444	0.504	0.493	0.508	0.457	0.437
Observed	0.250	0.233	0.188	0.273	0.308	0.277	0.222	0.129	0.244	0.236
Mean expected	0.455	0.423	0.341	0.336	0.326	0.261	0.286	0.278	0.236	
Mean observed	0.455	0.405	0.403	0.279	0.308	0.207	0.220	0.177	0.188	

Table 5 Power of discrimination (PD) values

PD	Cameroon	Rwanda	Kenya	Egypt	Oman	Galicia	Basque	China	Taiwan	Overall
11q22	0.7961	0.7309	0.7517	0.6585	0.6813	0.6096	0.5573	0.5030	0.2461	0.7323
K113	0.2971	0.5140	0.6027	0.2006	0.1283	0.0797	0.0761	0.3657	0.4292	0.3737
K115	0.5086	0.5140	0.5097	0.3815	0.4107	0.2350	0.1795	0.1316	0.1480	0.3697
Z80898a	0.2971	0.5766	0.5794	0.6202	0.5897	0.6187	0.6246	0.6247	0.5971	0.6128
Combined	0.9505	0.9731	0.9797	0.9359	0.9328	0.8952	0.8740	0.8973	0.8523	0.9591

Table 6 G scores and probability values for all pairwise combinations involving the nine populations analyzed

Population	Oman	Cameroon	Rwanda	Kenya	Basque	Galicia	Egypt	Taiwan	China
Oman		24.748	29.062	45.759	22.085	22.152	5.703	145.965	75.698
Cameroon	0.000		18.053	18.730	54.687	53.176	31.351	112.640	62.763
Rwanda	0.000	0.028		10.215	66.234	71.994	23.281	171.677	94.441
Kenya	0.000	0.026	0.255		83.456	89.913	42.719	172.461	100.222
Basque	0.005	0.000	0.000	0.000		22.401	15.603	169.122	96.893
Galicia	0.000	0.000	0.000	0.000	0.003		24.321	93.688	44.442
Egypt	0.685	0.000	0.002	0.000	0.079	0.001		140.800	74.592
Taiwan	0.000	0.000	0.000	0.000	0.000	0.000	0.000		17.503
China	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.030	

Table 7 Components of genetic variation across all populations

Genetic System	Hs (Intra-populations)					Gst (Inter-population)				
	WP ^a	SS ^b	NA/A ^c	EA ^d	EU ^e	WP ^a	SS ^b	NA/A ^c	EA ^d	EU ^e
HERV-K 11q22	0.458	0.590	0.516	0.224	0.434	0.211	0.016	0.003	0.020	0.121
HERV-K 113	0.191	0.326	0.088	0.244	0.040	0.105	0.067	0.003	0.003	< 0.000
HERV-K 115	0.210	0.344	0.242	0.074	0.113	0.052	< 0.000	0.001	< 0.000	0.002
HERV-K Z80898a	0.432	0.339	0.465	0.475	0.493	0.095	0.055	0.012	0.020	0.003
All loci	0.323	0.400	0.328	0.254	0.270	0.135	0.032	0.006	0.014	0.054

^aWhole set of populations.

^bsub-Saharan populations: Cameroon, Kenya, Rwanda.

^cNorth African/Arabian Peninsula populations: Egypt, Oman.

^dEast Asian populations: China, Taiwan.

^eEuropean populations: Galicia, Basques

into sub-Saharan, North African/Arabic Peninsula, East Asia and European groups of populations (Table 7). At the intra-population level, the sub-Saharan group of populations possesses the highest overall H_s value (0.400) while the East Asians have the lowest (0.254). Except for Z80898a, the sub-Saharan group exhibits the highest H_s values. In terms of inter-population variation, the worldwide set of populations provided the highest G_{sts} , both for each individual locus and overall, while the North African/Arabic Peninsula generated the overall lowest values.

The sub-Saharan and the East Asian populations cluster at opposite ends of the ML tree (Fig. 1) while the Middle Eastern groups exhibit an intermediate position between the two. A hypothetical ancestral population set at zero for all insertional alleles at the four loci segregate with the sub-Saharan groups. Except for the split between Kenya and Rwanda, the bifurcations' statistical support are all above 50% with one-half of the bootstrap values above 90%. The hypothetical outgroup clustered with the Cameroon population with a bootstrap value of 100%. The PC (Fig. 2) analysis exhibits three primary clusters: African, Asian and European/Middle-Eastern. The vertical axis separates the Caucasian populations, located in the lower left portion of the plot, from the sub-Saharan and East Asian groups. The East Asian populations are located in the upper left corner of the graph. Within the sub-Saharan cluster, the Kenyans from East Africa are located at the extreme upper right corner while the West African population from Cameroon plots at the opposite end. Rwanda, geographically located in between the East African population from Kenya and the West Africans from Cameroon, maps intermediate in the sub-Saharan sector.

In the Centroid analysis (Fig. 3) the sub-Saharan populations plot above the theoretical regression line.

This is indicative of higher than average gene flow and/or effective population size. The other groups from northern Africa, Asia, and Europe segregate below the regression line, diagnostic of lower than average gene flow and/or effective population size.

Two clades are clearly evident in the phylogenetic analysis of PRVIs and fixed HERV-K insertions

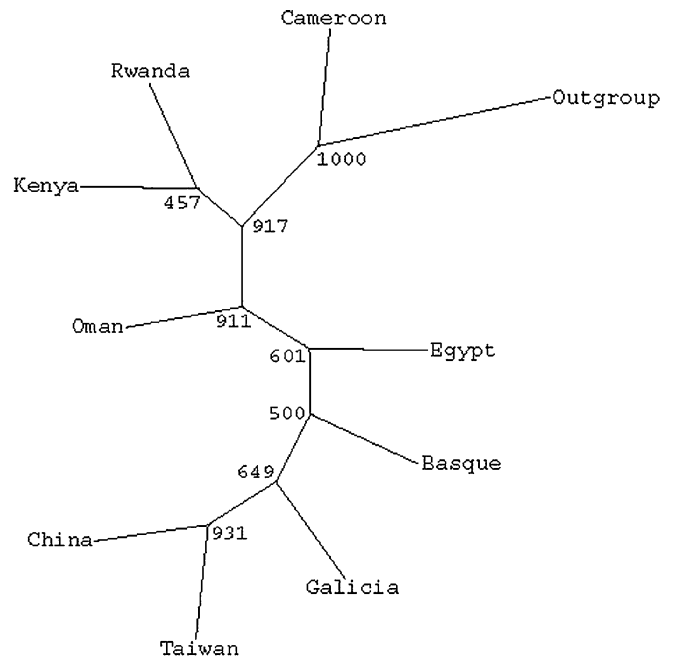
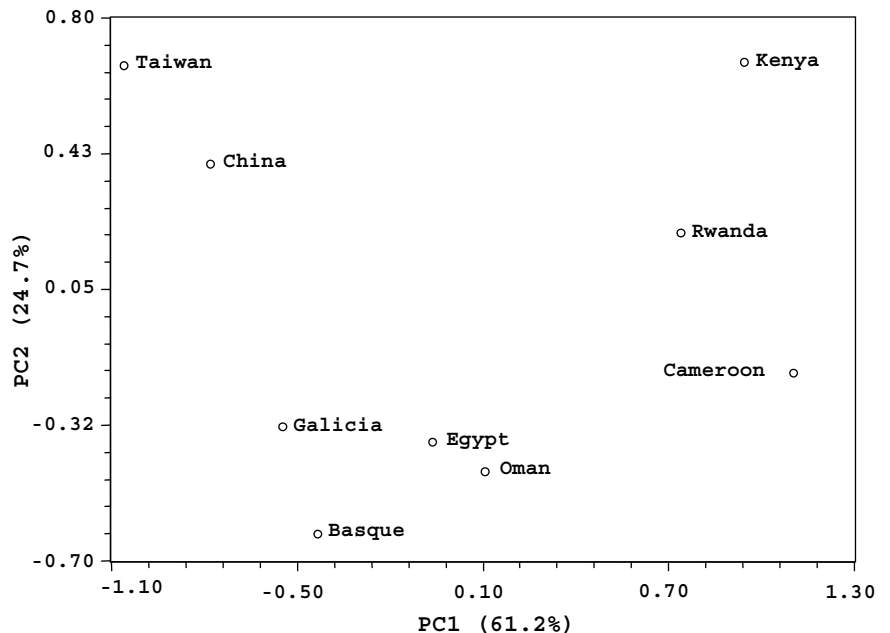


Fig. 1 Maximum likelihood (ML) tree derived from the allelic frequencies of the four human endogenous retrovirus K (HERV-K) loci in the nine populations analyzed in this study using the distance of Nei and Roychoudhury (1974). Numbers indicate bootstrap support for interior tree branches

Fig. 2 Principal component (PC) analysis of the nine populations analyzed in this study based upon frequencies for the four polymorphic retroviral insertions (PRVIs)



(Fig. 4). One clade in the tree contains the fixed 5' and 3' LTRs of the intact provirus HERV-K10, all but one of the PRVIs, and the fixed solo LTR C. The other clade encompasses four fixed solo LTR elements (D, E, F and G) as well as the Z80898a PRVI in the form of a solo-LTR. Fixed solo-LTR B is positioned basally to both groups. Nine out of 14 bifurcations exhibit bootstrap values above 50% and 7 out of 14 nodes are at or above 90%. The split separating the two main groups of LTR sequences is supported by a bootstrap value of 100%.

To assess the genetic distance relationship of the PRVIs examined in this study and other human-specific retroviral insertions, a PC analysis was conducted (Fig. 5). The PC plot illustrates a dramatic distinction in the spread of most fixed solo-LTRs in contrast to younger PRVIs. The segregation of the majority of the fixed solo-LTRs over a considerable area of the graph is diagnostic of marked genetic differentiation. Fixed solo-LTRs D, F, and G cluster in the upper left-hand corner while fixed solo-LTR E and polymorphic Z80898a segregate loosely at the lower left-hand corner. On the other hand, the group composed mostly of polymorphic HERV-K insertions is located in the right-hand side of the graph in a tight clump. An expansion of this compact cluster reveals (see enlarged view in Fig. 5) the segregation of the polymorphic loci into a compressed grouping that also includes the fixed HERV-K10 provirus. Fixed solo-LTRs B and C barely plot away from the intimate PRVI cluster.

Discussion

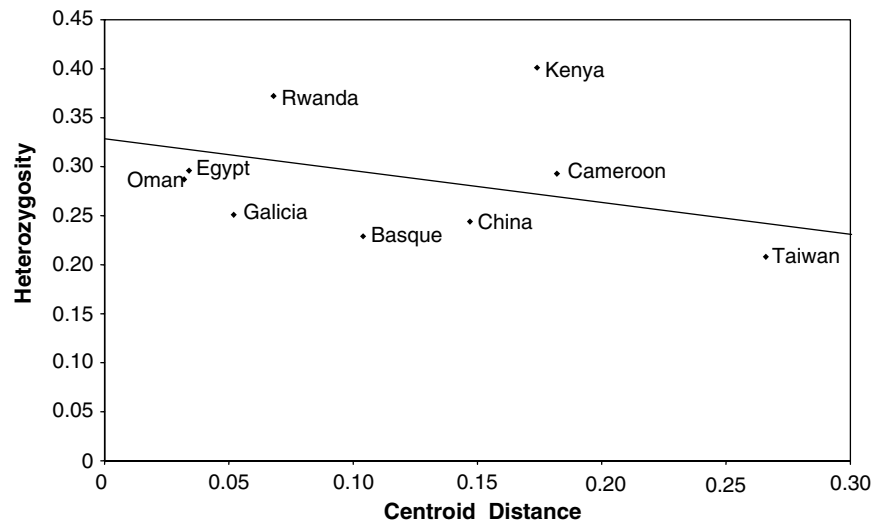
In the present study, four members of the HERV-K subfamily of retroviral insertions are examined as a novel genetic marker system for the study of recent human evolution. The HERV-K loci investigated provide patterns of genetic relatedness that are consistent not only with other traditional marker systems but also with reported historical and biogeographical data (Nei

and Roychoudhury 1974; Cavalli-Sforza et al. 1994; Hammer et al. 2000; King and Underhill 2002; Cinnioğlu et al. 2004; Semino et al. 2004).

Similar to previously published results that separates East Asians and Europeans from sub-Saharan Africans (Nei and Roychoudhury 1974; Cavalli-Sforza et al. 1994), the ML tree based on HERV-K data (Fig. 1) illustrates closer affinity of the Europeans to the East Asian populations. This dendrogram also segregates these three major biogeographical groups with high bootstrap support. At intermediate positions on the phylogram, populations that are geographically closer group together. Noteworthy is the clustering of the sub-Saharan African populations with a hypothetical ancestral group set homozygous for the lack of the insertion allele at all loci. This relationship supports the Out of Africa hypothesis of human origins and validates the usefulness of HERV-K insertion polymorphisms for phylogenetic studies. The segregation of a hypothetical ancestral population, lacking insertional elements at all loci, with sub-Saharan African groups has also been reported for polymorphic *Alu* insertions (PAIs) (Novick et al. 1998; Antunez de Mayolo et al. 2002). The African origin of modern man is further supported by the higher genetic diversity of the sub-Saharan populations reflected in the superior effective population size and/or gene flow observed in the HERV-K Centroid analysis (Fig. 3).

The PC analysis (Fig. 2) also exhibits clusters that are consistent with known genetic associations among human populations. As in the ML phylogram, the PC test places the Middle Eastern and European populations in intermediate positions in between the two extreme clusters, the East Asians and the sub-Saharan Africans. The greater diversity among the sub-Saharan populations in comparison to the Caucasian groups from the Middle East and Europe is also reflected in the lack of compactness of the former cluster. This higher level of diversity within the sub-Saharan African populations is further evident in a number of population genetic parameters discussed below.

Fig. 3 Plot of heterozygosity vs distance from the Centroid. The line represents the linear regression based on the expected heterozygosity values. Position above the line is indicative of greater than expected gene flow and/or effective population size



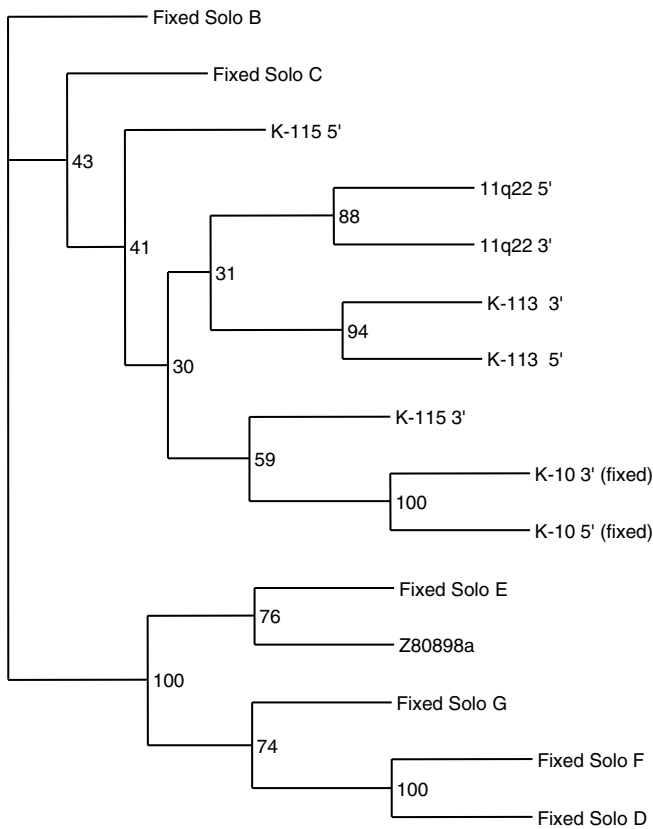
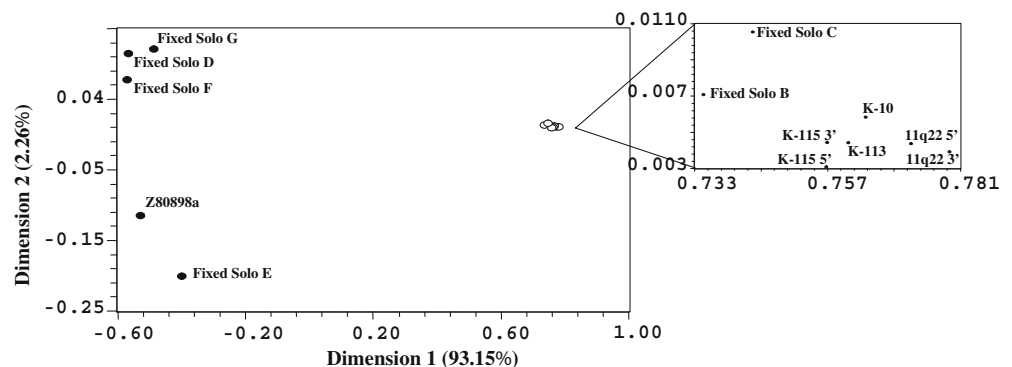


Fig. 4 The neighbor joining (NJ) tree of HERV-K elements was generated utilizing Gregorius' absolute genetic distance (1984). To explore the phylogenetic affinities among the PRVIs examined in this study as well as their relation to fixed HERV-K loci, a NJ analysis was performed based on a 968 bp long terminal repeat (LTR) sequence. The fixed loci encompass members of Cluster 9 as reported by Medstrand and Mager (1998). The following list includes the fixed solo LTRs investigated in this analysis and their GenBank accession numbers: Fixed Solo B (AC002508), C (U73641a), D (L47334), E (Z84493), F (AC002350), and G (U47924). As designated in Medstrand and Mager (1998), a letter was added at the end of the accession number of fixed solo-LTR C (e.g., U73641a) to indicate the presence of more than one LTR within the entry

In terms of the overall observed heterozygosity values, a trend is observed in which the sub-Saharan group of populations exhibit the highest levels and a decline is evident eastward and westward toward East Asia and Europe, respectively. This cline confirms the known

Fig. 5 A PC analysis based on distances from Clustal-X was performed in NTSYpc program version 2.02i utilizing the LTR sequences examined in Fig. 4. In this study, the HERV-K10 and HERV-K113 insertions are represented only by one LTR since the 5' and 3' LTRs possess identical sequences



higher levels of diversity in sub-Saharan Africa (Cavalli-Sforza 1998) and the expected reduction in variability as modern humans migrated to all corners of the world. It is also supportive of the Out of Africa hypothesis. These results are also reflected in the higher intra-population variation of sub-Saharan populations (Table 7). For example, gradients of decreasing H_s values are seen bidirectionally from Africa toward Europe and East Asia. These demic flows are reflected in the individual (locus-specific) intra-population variation values, as well as in the combined levels of all loci. Similarly in the Centroid plot, the sub-Saharan African populations exhibit higher than average gene flow and/or effective population size. Since higher values of gene flow and effective population size contribute to elevated levels of diversity and heterozygosity, all of these analyses corroborate each other and validate this novel marker system.

Of particular interest are the considerable discrepancies between observed and expected heterozygosities observed for the Z80898a locus. Except for Cameroon, the other eight worldwide populations exhibit lower than expected heterozygosity values, in some instances 75% less. In connection with these observations, Hardy-Weinberg equilibrium expectations are violated in six of the nine populations examined even after Bonferroni corrections. In these cases, the unrealized Hardy-Weinberg premises are likely to be the result of heterozygous deficit. It is possible that heterozygotes may be under negative selection pandemically. Previous studies have proposed the possibility that the Z80898a locus is under bidirectional selection, possibly due to its close proximity to the multi-histocompatibility complex (Medstrand and Mager 1998).

In terms of the PD afforded by these polymorphic HERV-K insertions, it is quite remarkable that only four loci are capable of generating an overall PD value of 0.9591. With the discovery of additional PRVIs, PD values will augment and provide an excellent genetic system for individual diagnosis and identification. In other words, these polymorphic retroviral markers have the potential of complementing other more traditional medical and forensic systems.

In order to examine the genetic relationships of the four PRVI elements to a number of human-specific fixed HERV-K insertions, a NJ phylogram was generated (Fig. 4). The clustering of the polymorphic loci together

as well as their segregation from the fixed solo-LTRs confirms the greater genetic similarity among the most recent polymorphic insertions in contrast to the older fixed elements. Although a fixed provirus (HERV-K10) and two fixed solo-LTRs (B and C) cluster with the PRVIs, and a PRVI segregates with the fixed solo-LTRs, there is an evident and basal (within the tree) separation of these two categories of HERV-K insertions. The location of Z80898a within the fixed solo-LTRs may be indicative of an older than usual provirus and/or a greater accumulation of genetic changes at this locus. The notion of Z80898a representing a locus under balancing selection may help explain the persistence of the retroviral insertional polymorphism in spite of its mature age. Since the sequence divergence between the 5' and 3' LTRs starts just after the insertional event, it is reassuring to observe the segregation of homologous LTR pairs as sister sequences in most of the proviral elements. From a practical perspective, the sequence relationships afforded by this NJ analysis may allow us to pinpoint diagnostic nucleotide differences capable of identifying recent retroviral insertions. Additional PRVIs would add to the usefulness of this novel genetic marker system as a diagnostic tool for individual identification as well as for phylogenetic analysis.

As reflected in the NJ dendrogram, the PC analysis illustrates a clear separation between the fixed solo-LTR loci and the polymorphic HERV-K insertions. The PC plot also illustrates large genetic distances between the predominantly fixed solo-LTR group and the close aggregation consisting mainly of polymorphic HERV-K insertions. Furthermore, D, E, F and G fixed solo-LTRs as well as Z80898a exhibit considerable genetic distance from each other. Although fixed solo-LTRs D, G and F segregate somewhat proximal to each other, fixed solo-LTR E and Z80898a are distant to each other and to the former solo-LTR trio. The distant geography of these five loci most likely is indicative of substantial divergence time in comparison to the small separation observed among most of the PRVIs. The tight conglomeration displayed by the polymorphic insertions most probably is the result of limited divergence time among them. These PRVIs most probably represent more modern insertional events. Although the fixed solo-LTRs B and C as well as the HERV-K10 monomorphic provirus plot close to the PRVIs, they clearly segregate away from the intimate cluster of polymorphic insertional loci. The fixed solo-LTRs B and C, and HERV-K10 may represent insertional events somewhat older than the PRVIs.

References

- Antunez de Mayolo G, Antunez de Mayolo A, Antunez de Mayolo P, Papiha SS, Hammer M, Yunis JJ, Yunis EJ, Damodaran C, Martinez de Pancorbo M, Caeiro JL, Puzyrev VP, Herrera RJ (2002) Phylogenetics of worldwide human populations as determined by polymorphic Alu insertions. *Electrophoresis* 23:3346–3356
- Barbulescu M, Turner G, Seaman MI, Deinard AS, Kidd K, Lenz J (1999) Many human endogenous retrovirus K (HERV-K) proviruses are unique to humans. *Curr Biol* 9:861–868
- Black WC IV (1997) BIOSYS-2. A computer program for the analysis of allelic variation in genetics. Colorado State University, Ft. Collins, CO
- Bock M, Stoye JP (2000) Endogenous retroviruses and the human germline. *Curr Opin Genet Dev* 10:651–655
- Brosius J (1999) RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. *Gene* 238:115–134
- Buzdin A, Ustyugova S, Khodosevich K, Mamedov I, Lebedev Y, Hunsmann G, Sverdlov E (2003) Human-specific subfamilies of HERV-K (HML-2) long terminal repeats: three master genes were active simultaneously during branching of hominoid lineages. *Genomics* 81:149–156
- Carmody G (1991) G-Test. Department of Biology, Carleton University, Ottawa
- Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The history and geography of human genes. Princeton University Press, New Jersey
- Chen CK, Burns ME, He W, Wensel TG, Baylor DA, Simon MI (2000) Slowed recovery of rod photoresponse in mice lacking the GTPase accelerating protein RGS9-1. *Nature* 403:557–560
- Cinnioglu C, King R, Kivisild T, Kalfoglu E, Atasoy S, Cavalleri GL, Lillie AS, Roseman CC, Lin AA, Prince K, Oefner PJ, Shen P, Semino O, Cavalli-Sforza LL, Underhill PA (2004) Excavating Y-chromosome haplotype strata in Anatolia. *Hum Genet* 114:127–48
- Desmarais D, Zhong Y, Chakraborty R, Perreault C, Busque L (1998) Development of a highly polymorphic STR marker for identity testing purposes at the human androgen receptor gene (HUMARA). *J Forensic Sci* 43:1046–1049
- Dunn CA, Medstrand P, Mager DL (2003) An endogenous retroviral long terminal repeat is the dominant promoter for human β 1,3-galactosyltransferase 5 in the colon. *Proc Natl Acad Sci USA* 100:12841–12846
- Felsenstein J (1989) PHYLIP: phylogeny inference package. *Cladistics* 5:164–166
- Gregorius HR (1984) A unique genetic distance. *Biometrics* J 26:13–18
- Hammer MF, Redd AJ, Wood ET, Bonner MR, Jarjanazi H, Karafet T, Santachiara-Benerecetti S, Oppenheim A, Jobling MA, Jenkins T, Ostrer H, Bonne-Tamir B (2000) Jewish and Middle Eastern non-Jewish populations share a common pool of Y-chromosome biallelic haplotypes. *Proc Natl Acad Sci USA* 97:6769–6774
- Harpending HC, Ward RH (1982) Chemical systematics and human population. In: Nitecki M (ed) *Biochemical aspects of evolutionary biology*. University of Chicago Press, Chicago, pp 213–252
- Hughes JF, Coffin JM (2004) Human endogenous retrovirus K solo-LTR formation and insertional polymorphisms: implications for human and viral evolution. *Proc Natl Acad Sci USA* 101:1668–1672
- Johnson WE, Coffin JM (1999) Constructing primate phylogenies from ancient retroviral sequences. *Proc Natl Acad Sci USA* 96:10254–10260
- Jordan K, McDonald JF (2002) A biologically active family of human endogenous retroviruses evolved from an ancient inactive lineage. *Genome Lett* 1:1–5
- King R, Underhill PA (2002) Congruent distribution of Neolithic painted pottery and ceramic figurines with Y-chromosome lineages. *Antiquity* 76:707–714
- Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W et al (2001) International Human Genome Sequencing Consortium: Initial sequencing and analysis of the human genome. *Nature* 409:860–921
- Lavie L, Medstrand P, Schempp W, Meese E, Mayer J (2004) Human endogenous retrovirus family HERV-K (HML-5):

- status, evolution, and reconstruction of an ancient betaretrovirus in the human genome. *J Virol* 78:8788–8798
- Leib-Mosch C, Seifarth W (1996) Evolution and biological significance of human retroelements. *Virus Genes* 11:133–145
- Li CC (1976) *First course in population genetics*. Boxwood Press, Pacific Grove, CA
- Löwer R, Löwer J, Kurth R (1996) The viruses in all of us: characteristics and biological significance of human endogenous retrovirus sequences. *Proc Natl Acad Sci USA* 93:5177–5184
- Löwer R (1999) The pathogenic potential of endogenous retroviruses: facts and fantasies. *Trends Microbiol* 7:350–356
- Macfarlane C, Simmonds P (2004) Allelic variation of HERV-K(HML-2) endogenous retroviral elements in human populations. *J Mol Evol* 59:642–56
- Mamedov I, Lebedev Y, Hunsmann G, Khusnutdinova, Sverdlov E (2004) A rare event of insertion polymorphism of a HERV-K LTR in the human genome. *Genomics* 84:596–599
- Mariani-Costantini R, Horn TM, Callahan R (1989) Ancestry of a human endogenous retrovirus family. *J Virol* 63:4982–4985
- Mayer J, Stuhr T, Reus K, Maldener E, Kitova M, Asmus F, Meese E (2005) Haplotype analysis of the human endogenous retrovirus locus HERV-K(HML-2.HOM) and its evolutionary implications. *J Mol Evol* 61:706–715
- McDonald J (1995) Transposable elements: possible catalysts of organismic evolution. *Trends Ecol Evol* 10:23–26
- Medstrand P, Mager D (1998) Human-specific integrations of the HERV-K endogenous retrovirus family. *J Virol* 72:9782–9787
- Nei M, Roychoudhury AK (1974) Genetic variation within and between the three major races of man, Caucasoids, Negroids, and Mongoloids. *Am J Hum Genet* 26:421–443
- Novick GE, Novick CC, Yunis E, Yunis J, Antunez de Mayolo P, Scheer WD, Deininger PL, Stoneking M, Batzer MA, Herrera RJ (1998) On the oriental origin of Native American populations. *Hum Biol* 70:23–39
- Ota T (1993) *DISPAN genetic distance and phylogenetic analysis*. Pennsylvania State University
- Patience C, Wilkinson DA, Weiss RA (1997) Our retroviral heritage. *Trends Genet* 13:116–120
- Reus K, Mayer J, Sauter M, Zischler H, Muller-Lantzsch N, Meese E (2001) HERV-K (OLD): ancestor sequences of the human endogenous retrovirus family HERV-K (HML-2). *J Virol* 75:8917–8926
- Rohlf FJ (1993) *NTSYS-pc. Numerical taxonomy and multivariate analysis system*. Exeter Software, New York
- Sambrook J, Fritsch E, Maniatis T (1989) *Molecular cloning: a laboratory manual*, 2nd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY
- Semino O, Magri C, Benuzzi G, Lin AA, Al-Zahery N, Battaglia V, Maccioni L, Triantaphyllidis C, Shen P, Oefner PJ, Zhivotovsky LA, King R, Torroni A, Cavalli-Sforza LL, Underhill PA, Santachiara-Benerecetti AS (2004) Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: inferences on the neolithization of Europe and later migratory events in the Mediterranean area. *Am J Hum Genet* 74:1023–1034
- Smit AFA (1996) The origin of interspersed repeats in the human genome. *Curr Opin Genet Dev* 6:743–748
- Sverdlov ED (2000) Retroviruses and primate evolution. *Bioessays* 22:161–171
- Tristem M (2000) Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database. *J Virol* 74:3715–3730
- Turner G, Barbulescu M, Su M, Seaman MI, Kidd KK, Lenz J (2001) Insertional polymorphisms of full-length endogenous retroviruses in humans. *Curr Biol* 11:1531–1535
- Villareal LP (1997) On viruses, sex and motherhood. *J Virol* 71:859–865
- Von Sternberg RM, Novick GE, Gao GP, Herrera RJ (1992) Genome canalization: the coevolution of transposable and interspersed repetitive elements with single copy DNA. *Genetica* 86:215–246
- Wilkinson DA, Mager DL, Leong J-A (1994) Endogenous human retroviruses. In: Levy JA (ed) *The Retroviridae*, vol 3. Plenum, New York, pp 465–534