High Copy Number in Human Endogenous Retrovirus Families is Associated with Copying Mechanisms in Addition to Reinfection

Robert Belshaw, Aris Katzourakis,*[1] Jan Pačes,† Austin Burt,* and Michael Tristem**

*Department of Biological Sciences, Imperial College London, Silwood Park Campus, Ascot, United Kingdom; and
†Institute of Molecular Genetics, Academy of Sciences, Prague, Czech Republic

There are at least 31 families of human endogenous retroviruses (HERVs), each derived from an independent infection by an exogenous virus. Using evidence of purifying selection on HERV genes, we have shown previously that reinfection by replication-competent elements was the predominant mechanism of copying in some families. Here we analyze the evolution of 17 HERV families using d_N/d_S ratios and find a positive relationship between copy number and the use of additional copying mechanisms. All families with more than 200 elements have also used one or more of the following mechanisms: (1) complementation in *trans* (elements copied by other elements of the same family; HERV-H and ERV-9), (2) retrotransposition in *cis* (elements copying themselves) within germ-line cells (HERV-K(HML3)), and (3) being copied by non-HERV machinery (HERV-W). We discuss why these other mechanisms are rare in most families and suggest why complementation in *trans* is significant only in the larger families.

Introduction

Endogenous retroviruses (ERVs) are the proviral form of exogenous retroviruses that have become integrated into the germ line of the host (Boeke and Stoye 1997). The human genome contains 98,000 such ERVs (J. Pačes, Pavlíček, and V. Pačes 2002), and together with the 158,000 mammalian apparent long terminal repeat (LTR) retrotransposons (MaLRs), they make up 8% of our genome (IHGSC 2001).

Typically a human endogenous retrovirus (HERV) element consists of an internal region of three genes (*gag*, *pol*, and *env*) flanked by two sequences known as LTRs, which are identical at the time of integration and are essential for replication. Katzourakis and Tristem (2005) defined 31 HERV families, each of which is considered to be a clade derived from a single infection of the human germ line (Tristem 2000). Most HERV elements integrated into the genome tens of millions of years ago and have accumulated numerous stop codons and frameshift mutations or have undergone recombination between their LTRs, leading to the loss of the entire internal region and leaving only a solo LTR (Stoye 2001). All families except HERV-K(HML2) have long ceased proliferation (IHGSC 2001), and no active HERVs are known; thus, the copying mechanisms by which they proliferated can only be inferred indirectly.

We have examined previously the evolution of HERV-K(HML2) and several other families and found strong evidence of past purifying selection acting on the *env* gene (Belshaw et al. 2004), which is necessary only for movement between host cells. From this we inferred that most copying was via the reinfection of germ-line cells by replication-competent elements. To what extent this involved infectious transfer between host individuals or was simply the movement between cells of the same individual has yet to be determined. Acquisition of novel endogenous elements via reinfection has been demonstrated experimentally in mice, where endogenous elements can copy themselves into the germ line of offspring derived from transplanted and virus-free ovaries via infection from the host mother (Boeke and Stoye 1997). Here, we test the predominant role of reinfection by analyzing 17 HERV families, using a significantly reduced rate of nonsynonymous (d_N) compared to synonymous (d_S) nucleotide substitution as evidence of past purifying selection (Li 1997).

We find that all 13 families with a copy number below 200 have a low d_N/d_S ratio in the *env* gene (table 1). In each case, the ratio is significantly below 1 except for HERV-R, where $P = 0.053$. Reinfection of germ-line cells by replication-competent elements thus appears to be the predominant copying method in HERVs. However, the four families with a copy number above 200 (which are phylogenetically unrelated; Katzourakis and Tristem 2005) all show evidence of other mechanisms (see below). A binomial simulation (repeatedly taking the first four items from a shuffled list representing these four families and the 13 reinfecting families) shows that this is extremely unlikely to have occurred by chance ($P < 0.001$). There is also a significant correlation between copy number and the *env* d_N/d_S ratio (fig. 1), which we use as an estimate of the relative importance of reinfection by replication-competent elements (Spearman rank correlation; $\rho = 0.57$; $P = 0.01$).

The largest family, HERV-H, is dominated by a single subclade of elements that share large inactivating deletions and which is nested within a smaller paraphyletic grade of more intact elements (fig. 2). Elements in the small intact grade have a low *env* d_N/d_S ratio of 0.23 and appear to have been reinfecting as in the 13 smaller families. In contrast, elements in the deleted subclade appear to have been copied by proteins derived from the intact elements, a process called complementation in *trans*, as suggested by Mager and Freeman (1995) (see also the Supplementary Material online). We also find evidence of complementation in *trans* in the ERV9 family, which has a high d_N/d_S ratio both for *env* (0.79) and *pol* (0.57), showing a marked relaxation of purifying selection on both genes (although we did not find large shared deletions in this family). The rarity of complementation in *trans* among HERV families is surprising given that retroviral replication involves the obligate copackaging of two viral mRNAs within the same viral particle. We speculate that

[1] Present address: Department of Zoology, University of Oxford, Oxford, United Kingdom.

Key words: human, endogenous, retrovirus, infection, retrotransposition, complementation.

E-mail: r.belshaw@imperial.ac.uk.

Mol. Biol. Evol. 22(4):814–817. 2005
doi:10.1093/molbev/msi088
Advance Access publication January 19, 2005

Table 1
Copy Number and *env* d_N/d_S Ratio in HERV Families

Family	Copy Number	*env* d_N/d_S Ratio
HERV-H	1,306	0.235***[a], 0.945[NS,b]
ERV-9	418	0.790[NS]
HERV-W	315	0.361*[,c]
HERV-K(HML3)	230	0.733*
HERV-E	181	0.215***
HERV-S	165	0.213***
HERV-HML5	163	0.400***
HERV-K(HML2)	121	0.265***
HERV-I	111	0.092***
HERV-HML6	109	0.348***
HERV-K(HML7)	108	0.069***
HERV-K(HML1)	96	0.074***
HERV-R	91	0.450[NS]
HERV-T	60	0.119***
HERV-F type b	57	0.379***
RRHERV-I	50	0.050***
HERV-XA	48	0.030***

NOTE.—The significance of the difference in likelihood when the *env* d_N/d_S ratio is fixed at 1 is also shown (* $P < 0.05$; *** $P < 0.001$; NS, not significant).

[a] From elements in the intact grade.

[b] From elements in the deleted subclade.

[c] Excluding LINE-copied elements.

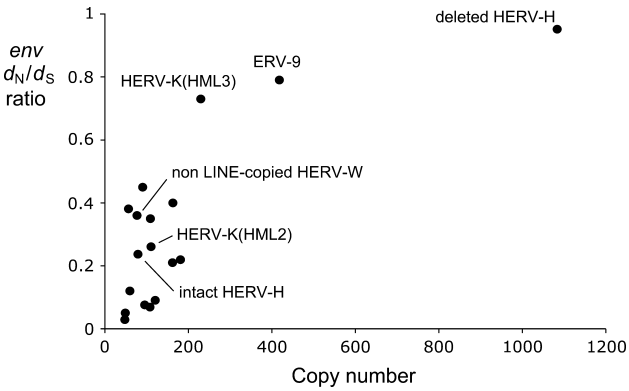

FIG. 1.—Relationship between *env* d_N/d_S ratio and copy number. We excluded LINE-copied HERV-W elements and divided HERV-H into two data points representing (1) the grade of intact elements and (2) the deleted subclade.

this is caused by a low probability that more than one element was expressed in the same cell at the same time. This would be consistent with complementation in *trans* being found only in the larger families because chance increases in copy number would have increased this probability and thus increased the opportunities for complementation in *trans*. Furthermore, if replication-competent elements were more deleterious to the host than nonautonomous ones (elements capable of being copied only by complementation in *trans*), a higher frequency of the latter would then have drifted to fixation.

We find evidence for a second additional mechanism in HERV-K(HML3), which has a very high *env* d_N/d_S ratio (0.73), while the *pol* d_N/d_S ratio (0.15) is as low as that in the predominantly reinfecting families such as HERV-K(HML2). We infer from this that many HERV-K(HML3) elements have used their own proteins to copy themselves intracellularly within germ-line cells—a process that would require a functional *gag* and *pol* but not *env* (called retrotransposition in *cis*). Perhaps the most likely explanation for the rarity of this mechanism among HERV families is that the act of budding through the host cell wall was essential for the formation of infectious viral particles, which is the case with many exogenous retroviruses (Swanstrom and Wills 1997). An alternative explanation is that germ-line cells may have been able to suppress HERV expression, but we consider this unlikely because the reported expression of HERVs tends to be in germ-line cells and early in development (Löwer 1999). We also infer that (1) there has been some reinfection because the *env* d_N/d_S ratio, although high, is significantly below 1 ($P < 0.05$) and (2) there has been some complementation in *trans* because approximately one-quarter of the elements share an inactivating deletion in *pol* (Mayer and Meese 2002; these elements are not included in our d_N/d_S analysis).

A third additional mechanism is known from HERV-W, two-thirds of whose elements have been copied by another type of retrotransposon called a LINE (long interspersed nuclear element; Pavlíček et al. 2002). Such LINE-copied elements lack promoter sequences (so cannot be transcribed further) and are dispersed through the phylogeny of the family (Costas 2002), showing that they have been derived from many different elements. This mechanism has been rare outside of the HERV-W family, but a plausible explanation for this is lacking. The remaining members of the family have a low *env* d_N/d_S ratio (0.36) that is significantly below 1 ($P < 0.05$), and the family therefore contained a core of reinfecting elements.

We have not analyzed some large groups of LTR elements in the human genome. The second largest HERV family is HERV-L, which may be over 70 Myr old (Bénit et al. 1999); the family lacks *env* and is thus assumed to have proliferated by copying within germ-line cells. Also, the abundant MaLRs are thought to be nonautonomous (Smit 1996; although it is not known if they are a natural group in the sense of the HERV families). Thus, it appears that reinfection by replication-competent elements has driven the evolution of most endogenous retrovirus lineages in our genome but was directly responsible only for a minority of the individual integrations that became fixed.

Methods

Our mining of HERVs is described in J. Pačes, Pavlíček, and V. Pačes (2002). For each family, we constructed a representative amino acid sequence for each gene by (1) finding open reading frames using getorf (Rice, Longden, and Bleasby 2000), (2) selecting the most representative ones using BlastAlign (Belshaw and Katzourakis, 2005) modified to use BlastP, and (3) confirming by blasting against GenBank. Nucleotide sequences were then aligned to the amino acid sequence using BlastAlignP (Belshaw and Katzourakis, 2005) and a Neighbor-Joining tree (HKY85 model) built using PAUP* (Swofford 1998). We calculated the d_N/d_S ratios on the internal branches of the tree using the ''two-ratio'' model in PAML (Yang 1997). This model allowed the largely neutral evolution represented by the terminal branches, when elements have become fixed and defective, to be ignored

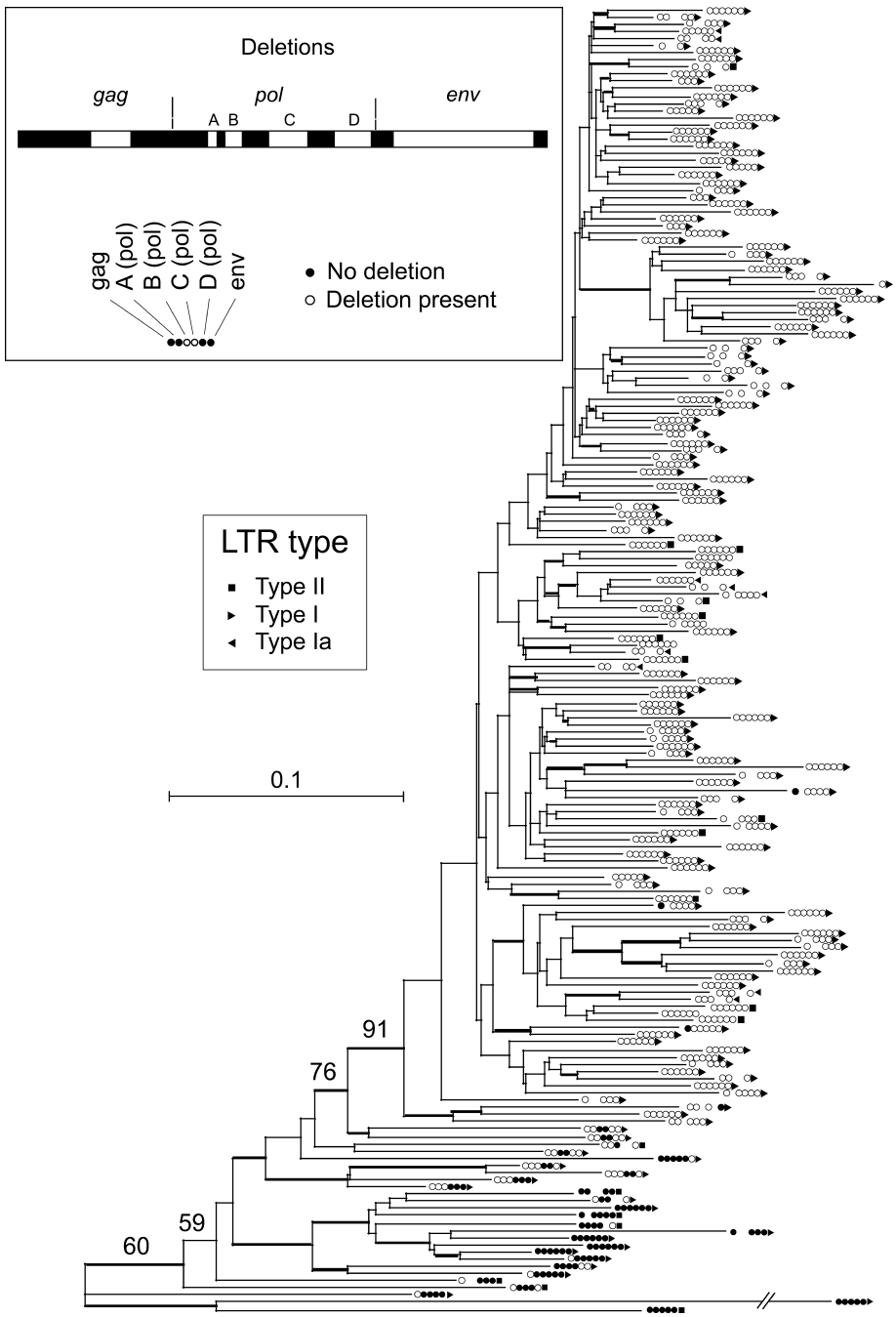

FIG. 2.—Maximum likelihood (ML) phylogeny of HERV-H. Bootstrap values (1,000 replicates) above 50% are shown on the spine of the tree, and branches are thickened if recovered in all most parsimonious trees (MPTs). The LTR type and the presence or absence of each deletion is indicated (the symbol is absent where this could not be determined). The scale bar shows nucleotide divergence. ML was implemented in PHYML (Guindon and Gascuel 2003; HKY + γ model) and maximum parsimony in PAUP* (10,000 random additions with Tree Bisection-Reconnection on one MPT).

(Belshaw et al. 2004). We took values significantly below 1 ($P < 0.05$) as showing past purifying selection. Significance was measured by finding the likelihood when the internal ratio was fixed at 1 and comparing twice the difference in log.likelihood to the χ^2 distribution with one degree of freedom (Yang 1998). To improve accuracy we ignored both old and small HERV families and excluded solo LTRs from the copy number (hence, we assume that rates of recombinational deletion do not vary markedly between families). Further details are in the Supplementary Material online.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online www.molbiolevol.org.

Acknowledgments

This work was funded by the Wellcome Trust. A.K. was in receipt of a Natural Environment Research Council Studentship and subsequently a Medical Research Council Fellowship, and J.P. was funded by the Centre for Integrated Genomics Grant LN00A079. We also thank Vini Pereira for assistance with the programing and discussion of ideas.

Literature Cited

Belshaw, R., and A. Katzourakis. 2005. *BlastAlign*: a program that uses *blast* to align problematic nucleotide sequences. Bioinformatics **21**:122–123.

Belshaw, R., V. Pereira, A. Katzourakis, G. Talbot, J. Pačes, A. Burt, and M. Tristem. 2004. Long-term re-infection of the human genome by endogenous retroviruses. Proc. Natl. Acad. Sci. USA **101**:4894–4899.

Bénit, L., J. B. Lallemand, J. F. Casella, H. Philippe, and T. Heidmann. 1999. ERV-L elements: a family of endogenous retrovirus-like elements active throughout the evolution of mammals. J. Virol. **73**:3301–3308.

Boeke, J. D., and J. P. Stoye. 1997. Retrotransposons, endogenous retroviruses, and the evolution of retroelements. Pp. 343–435 *in* J. M. Coffin, S. H. Hughes, and H. E. Varmus, eds. Retroviruses. Cold Spring Harbor Laboratory Press, Plainview, N.Y.

Costas, J. 2002. Characterization of the intragenomic spread of the human endogenous retrovirus family HERV-W. Mol. Biol. Evol. **19**:526–533.

Guindon, S., and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. **52**:696–704.

[IHGSC] International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. Nature **409**:860–921.

Katzourakis, A., and M. Tristem. 2005. Phylogeny of human endogenous and exogenous retroviruses. Pp. 186–203 *in* E. D. Sverdlov, ed. Retroviruses and primate genome evolution. Landes Bioscience, Georgetown, Tex.

Li, W.-H. 1997. Molecular evolution. Sinauer Associates, Sunderland, Mass.

Löwer, R. 1999. The pathogenic potential of endogenous retroviruses: facts and fantasies. Trends Microbiol. **7**:350–356.

Mager, D. L., and J. D. Freeman. 1995. HERV-H endogenous retroviruses: presence in the New World branch but amplification in the Old World primate lineage. Virology **213**: 395–404.

Mayer, J., and E. U. Meese. 2002. The human endogenous retrovirus family HERV-K(HML-3). Genomics **80**:331–343.

Pačes, J., A. Pavlíček, and V. Pačes. 2002. HERVd: database of human endogenous retroviruses. Nucleic Acids Res. **30**: 205–206.

Pavlíček, A., J. Pačes, D. Elleder, and J. Hejnar. 2002. Processed pseudogenes of human endogenous retroviruses generated by LINEs: their integration, stability and distribution. Genome Res. **12**:391–399.

Rice, P., I. S. Longden, and A. Bleasby. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. **16**:276–277.

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

Stoye, J. P. 2001. Endogenous retroviruses: still active after all these years? Curr. Biol. **11**:R914–R916.

Swanstrom, R., and J. W. Wills. 1997. Synthesis, assembly and processing of viral proteins. Pp. 263–334 *in* J. M. Coffin, S. H. Hughes, and H. E. Varmus, eds. Retroviruses. Cold Spring Harbor Laboratory Press, Plainview, N.Y.

Swofford, D. L. 1998. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, Mass.

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.

Yang, Z. H. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13**:555–556.

Yang, Z. H. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. **15**:568–573.

Lauren McIntyre, Associate Editor

Accepted January 3, 2005