Long-term reinfection of the human genome by endogenous retroviruses

Robert Belshaw[†‡], Vini Pereira[†], Aris Katzourakis[†], Gillian Talbot[†], Jan Pačes[§], Austin Burt[†], and Michael Tristem[†]

[†]Department of Biological Sciences, Imperial College at Silwood Park, Ascot, Berks SL5 7PY, United Kingdom; and [§]Institute of Molecular Genetics, Academy of Sciences, Prague 6, CZ-16637, Czech Republic

Edited by John M. Coffin, Tufts University School of Medicine, Boston, MA, and approved February 17, 2004 (received for review November 25, 2003)

Endogenous retrovirus (ERV) families are derived from their exogenous counterparts by means of a process of germ-line infection and proliferation within the host genome. Several families in the human and mouse genomes now consist of many hundreds of elements and, although several candidates have been proposed, the mechanism behind this proliferation has remained uncertain. To investigate this mechanism, we reconstructed the ratio of nonsynonymous to synonymous changes and the acquisition of stop codons during the evolution of the human ERV family HERV-K(HML2). We show that all genes, including the *env* gene, which is necessary only for movement between cells, have been under continuous purifying selection. This finding strongly suggests that the proliferation of this family has been almost entirely due to germ-line reinfection, rather than retrotransposition in *cis* or complementation in *trans*, and that an infectious pool of endogenous retroviruses has persisted within the primate lineage throughout the past 30 million years. Because many elements within this pool would have been unfixed, it is possible that the HERV-K(HML2) family still contains infectious elements at present, despite their apparent absence in the human genome sequence. Analysis of the *env* gene of eight other HERV families indicated that reinfection is likely to be the most common mechanism by which endogenous retroviruses proliferate in their hosts.

Endogenous retroviruses (ERVs) represent the proviral phase of exogenous retroviruses that have integrated into the germ line of their host (1). They typically consist of an internal region with three genes (*gag*, *pol*, and *env*) plus two flanking, noncoding LTRs, which are identical at the time of integration. Human ERVs (HERVs) comprise ≈5–8% of the human genome (2), with 98,000 elements and fragments (3), but phylogenetic analysis of conserved regions within their *pol* and *env* genes indicates that they form only a small number of clades among nonhuman exogenous and endogenous retroviruses (4–6). Thus, there appears to have been a huge proliferation of elements derived from only a few initial germ-line invasions by exogenous viruses. Over time, replication-competent ERVs accumulate in-frame stop codons and frame-shift mutations as a result of host DNA replication and, within the human genome sequence, these processes have led to the inactivation of almost every element (2). Another mechanism by which ERVs are inactivated is via recombinational deletion between the two viral LTRs, which removes the internal region leaving a solo LTR structure (7). Solo LTRs are typically 10–100 times more numerous than their more intact, undeleted counterparts (7).

Surprisingly, although the processes that inactivate ERVs are known, the mechanism by which HERVs have increased in copy number is only poorly understood, though several candidate mechanisms have been proposed (1). ERVs could undergo retrotransposition within germ-line cells, and do this via two routes: either in *cis*, where the virus uses its own encoded proteins for mobilization [the predominant method in long interspersed nuclear elements (LINEs); ref. 8], or by complementation in *trans*, where the proteins are supplied by another endogenous or exogenous virus within the same cell. Retrotransposition in *cis* does not require an intact *env* gene (which is necessary only for movement outside the cell), whereas complementation in *trans* does not require the virus to have any functional genes (merely requiring a promoter and other motifs for expression and packaging of the viral RNA). ERVs could also increase in copy number by reinfection. Reinfection can occur between germ-line cells, or by infection of germ-line cells by viruses originating in somatic cells. Thus, it does not necessarily require the viruses to be passed between different individuals in a population. Studies on murine leukemia virus proviruses in mice (1) and *gypsy* retroelements in *Drosophila* (9) show that new elements can integrate into the germ line by extracellular infection. All these mechanisms can be expected to have left different patterns in the nucleotides of existing ERVs.

Within humans, the most recently active ERVs are members of the HERV-K (HML2) family (10, 11). This family first integrated into the genome of the common ancestor of humans and Old World monkeys at least 30 million years ago, and it contains >12 elements that have integrated since the divergence of humans and chimpanzees, as well as at least two that are polymorphic among humans (12–14). This recent activity makes this family ideal for distinguishing between the alternative mechanisms of proliferation.

Here we show that the HERV-K (HML2) family has increased in copy number predominantly via reinfection, and that the family has probably retained replication-competent and infectious members for >30 million years. We also present evidence for persistent reinfection by other ERV families within the human genome, and suggest that endogenous retrovirus families are often capable of extremely long periods of smoldering infection.

Materials and Methods

Mining. We used REPEATMASKER version 20020505 (http://ftp.genome.washington.edu/RM/RepeatMasker.html) and WU BLAST 2.0 (http://blast.wustl.edu), to search through build 31 of the human genome and identify elements homologous to HERV-K10 (GenBank accession no. M14123). PERL scripts then parsed the output, defragmented, and recovered elements that had both LTRs (these and other scripts used are available on request). Many of the HERV-K(HML2) elements, and their insertion sites, were already known from humans and other primates (13–15). By comparing flanking sequences using the NCBI BLAST genome web site, we (*i*) inserted into our alignment three elements (3q27, K103 and K113), which we found to be absent from build 31 or were represented only by solo LTRs (refs. 13–15; except for these elements we use our own nomenclature); (*ii*) identified those elements known to have orthologs in other primates; and (*iii*) confirmed segmental duplications, which are indicated by the presence of sister clades in the LTR

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: ERV, endogenous retrovirus; HERV, human ERV; MP, maximum parsimony.

[‡]To whom correspondence should be addressed. E-mail: r.belshaw@imperial.ac.uk.

tree: one composed only of 5′ LTRs and the other composed of the corresponding 3′ LTRs.

Alignment. Sequences were aligned with NCBI BLAST 2.2.5, by using HERV-K10 as the query and the flat query-anchored multiple alignment option. A PERL script was used to convert this alignment to nexus file format and exclude elements that were less than half the length of the final alignment. The alignment was then adjusted to maintain ORFs, which were determined from several GenBank-derived HERV-K(HML2) sequences. LTR sequences were aligned separately as above.

Phylogeny Estimation. Both Bayesian Markov Chain Monte Carlo inference and maximum parsimony (MP) were used to reconstruct the phylogeny of the internal region and of the LTRs. For the former we used MRBAYES 3.0B4 (16) (HKY85 + γ model). We ran four chains past their asymptotes, and then collected 5,000 trees (one per 100 generations), from which a majority-rule consensus phylogram was calculated. For MP, an unweighted heuristic search implemented in PAUP 4.0b10 (17) was performed. This used 10,000 random additions followed by branch swapping using tree bisection reconnection on a single tree. Trees were midpoint rooted because the low level of homology to other HERV families and exogenous viruses prevented accurate outgroup rooting.

Mapping the Acquisition of Stop Codons. We used maximum likelihood as implemented in MULTISTATE (18), finding the highest partial likelihood of states at internal nodes. The root was fixed as being without a stop codon because it is probable that the original, founding virus was fully functional. Where there was evidence of local optima, we took the highest likelihood from 20 runs. Several stop codons were estimated as being acquired on internal branches. To test the support for this, the unconstrained likelihood was compared to the likelihood when the acquisition of each stop codon was, in turn, constrained to a terminal branch. The likelihood differences were then summed and a likelihood ratio test was performed as described below.

Calculation of d_N/d_S Ratios. Nonsynonymous versus synonymous substitution ratios (d_N/d_S ratios) were calculated by using PAML 3.13 (19). To see whether there was a significant difference between the d_N/d_S ratio of internal compared to terminal branches, we estimated a single d_N/d_S ratio for the entire tree (the "one-ratio" model), and separate values for the internal and terminal (plus segmentally duplicated) branches (the "two-ratio" model). The significance of differences between the one-ratio and two-ratio models was then assessed by using a likelihood ratio test (20); twice the difference between the log likelihoods was compared to critical values on the χ^2 distribution with 1 df. The same likelihood ratio test was also used to compare unconstrained "two-ratio" models with those in which the d_N/d_S value for internal branches was fixed at a certain value. We also estimated a separate ratio for each branch of the phylogeny (the "free-ratio" model) for two purposes: first, to allow the values for each branch to be shown on the phylogeny, and second, to allow a second test of the difference in d_N/d_S ratios of internal compared to terminal branches. For this, the d_N/d_S ratio of each branch was treated as an independent observation in a Wilcoxon rank sum test.

Comparison with Exogenous Viruses and Other HERV Families. For comparison with exogenous viruses, we determined the d_N/d_S ratios in a clade of β-retroviruses, which are related to the HERV-K(HML2) family (6). This clade comprised simian retrovirus type 1 (GenBank accession nos. M11841 and U85505) and type 2 (GenBank accession nos. M16605, AF126467, and AF126468) plus Mason–Pfizer monkey virus (GenBank accession nos. AF033815 and M12349). We found *env* ORFs for other HERV families in the HERVd database (3), including representatives of all three retrovirus classes, by using GETORF (21). Nucleotide sequences were aligned to a composite ORF by using coordinates taken from a TBLASTN search. We then built a neighbor-joining tree by using PAUP (HKY + γ model) and analyzed the d_N/d_S ratios by using PAML as described above.

Tests for Recombination. In addition to examining the estimated phylogeny of the 5′ and 3′ LTRs, which can reveal nonhomologous crossing-over in the internal region, we investigated potential recombination events in the internal region directly by using several methods. For all these methods, we used a data set that excluded gaps. First, all taxa in which the proportion of gaps to nucleotides in the multiple alignment was >0.2 were removed, then all positions that contained gaps were removed. This left 32 taxa and an alignment spanning 4,279 bp. Initially we used PHYLPRO 0.8 (22), which uses a sliding window technique to identify potential breakpoints where there is a reduced correlation between two distance matrices calculated from either side. Possible recombinant and parental sequences were further investigated by using LARD 2.2 (23). LARD utilizes two putative parental and a single recombinant sequence to calculate both a single likelihood value and the sum of two likelihoods from either side of the putative breakpoint. In common with PHYLPRO, this method uses a sliding window for the analysis. The highest likelihood ratio was then compared to that observed in a null distribution of sequences (generated without recombination) by using SEQ-GEN 1.2.6 (24). We also searched for phylogenetic conflict by dividing the alignment into five contiguous data sets, each of 856 bp in length. A maximum parsimony analysis of each data set was then performed and a strict consensus of all trees was calculated. Any recombination of large fragments would be expected to result in an unresolved topology. Lastly, we investigated the overall impact of recombination by using the program PIST 1.0 (http://evolve.zoo.ox.ac.uk), which implements the informative sites test (25). This calculates the observed proportion of two-state parsimony informative characters to polymorphic characters, and compares this to a distribution derived from replicate data sets of clonally evolved sequences (1,000 replicates were used). The HKY + γ model, with parameter values estimated from a Bayesian tree, was used for this analysis. We also calculated an informative sites index (25), which provides a measure of the extent to which recombination has affected the phylogenetic pattern (we used 10 replicate data sets).

Results

Phylogeny Estimation. Phylogenetic analysis of the internal region and LTRs of the HERV-K (HML2) family revealed the presence of two lineages, one leading to a clade containing elements specific to humans, and another leading to a clade that has been copied by segmental duplication of the human genome (Fig. 1). Because past recombination events can distort the analysis of phylogenetic data, and evidence for recombination in the HERV-K(HML2) family has been presented (15, 26), we analyzed the impact of recombination on our phylogeny.

Overall, the internal region and LTR phylogenies were found to be largely congruent and, furthermore, there were only eight elements whose 5′ and 3′ LTRs did not form pairs on the LTR tree (excluding the segmentally duplicated clade) (Fig. 1). This nonpairing is not caused by recombination in the internal region, where we find evidence only for exchanges of small fragments of up to a few hundred base pairs in length (discussed below). Instead, consistent with previous results (15, 27), the nonpairing of LTRs is probably caused by gene conversion and recombination with solo LTRs, which are >10 times as common as paired ones (3). The absence of direct repeats in four of the six elements whose LTRs do not pair (the LTR/direct repeat boundary was

EVOLUTION

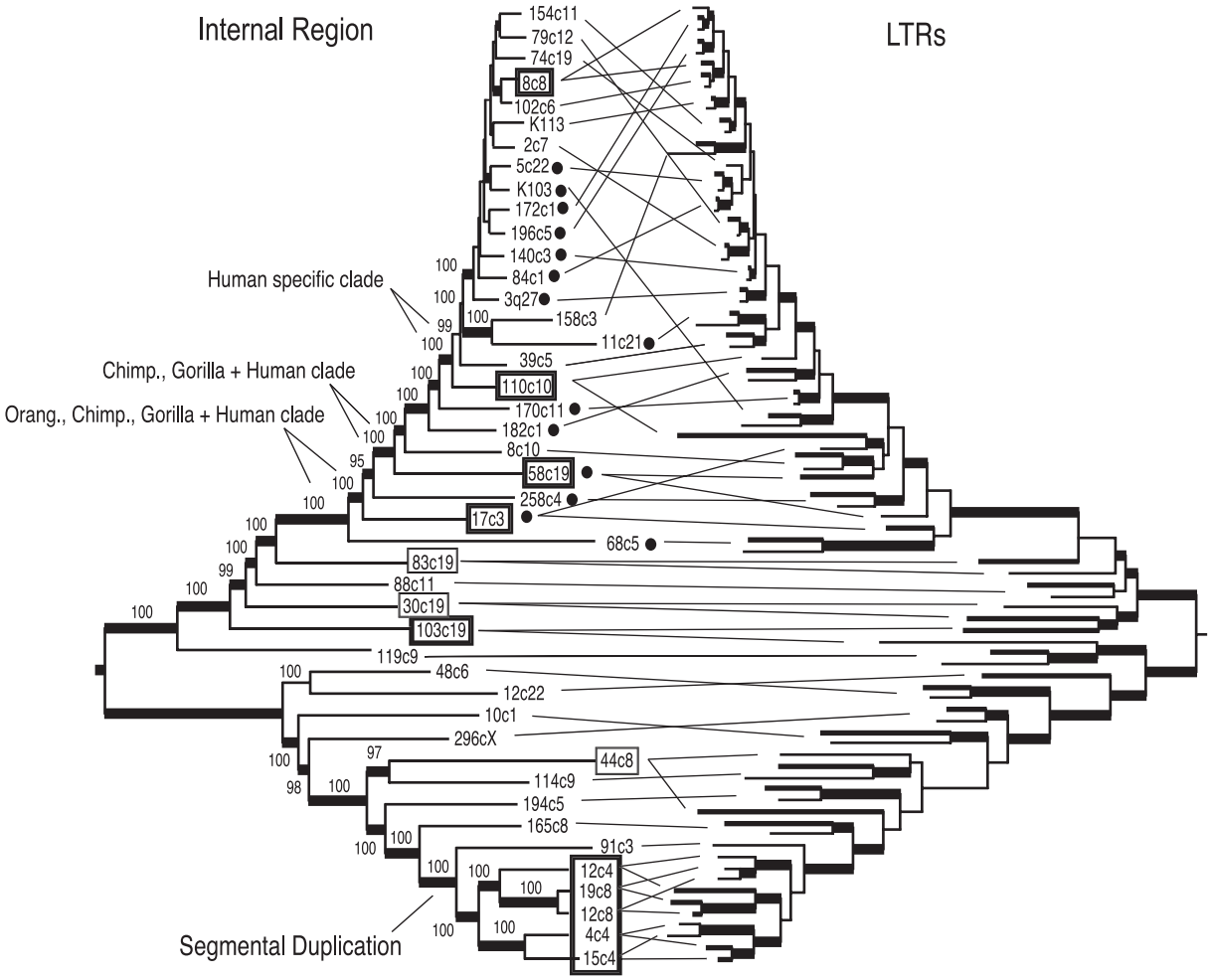

Fig. 1. Bayesian estimate of the phylogenies for the internal region and LTRs (those of the same element are linked by lines). Posterior probabilities for the internal region are shown above the branches except (for clarity) in the human-specific clade. Internal branches are thickened if they were also recovered in a strict consensus of the most parsimonious trees. Thick terminal branches in the LTR tree represent 5′ LTRs, and thin terminal branches represent 3′ LTRs. Boxes represent elements whose 5′ and 3′ LTRs do not cluster together, with a thick box showing that they are polyphyletic in both Bayesian and MP trees, and a thin box showing that they are paraphyletic in both, or monophyletic in one. Filled circles denote elements containing the 292-bp deletion spanning the *pol-env* junction [Type I HERV-K(HML2) subgroup], which is absent in all other elements [Type II HERV-K(HML2) subgroup]. The clade created by segmental duplication of the host genome is also indicated, as are the branches indicating probable changes in the distribution of elements among primates.

missing from the remaining two elements) suggests that recombination may be more common than gene conversion (unpublished data).

Within the internal region, the correlation profile of PHYLPRO (Fig. 2) is highly suggestive of recombination, with two recombinant groups (A and B), each composed of three elements with near identical profiles. Both of these two groups have multiple breakpoints (there are probably six breakpoints within group A), and all regions of recombination appear to be small, with a maximum size of only a few hundred base pairs. Increasing the size of the sliding window dramatically increased the correlation either side of the break point. We confirmed that recombination had occurred in one of these groups by likelihood analysis using LARD, identifying putative parental elements by visual inspection of the PHYLPRO correlation matrices. This found significant likelihood differences in the case of all three elements within group A and with breakpoints that corresponded exactly to the negative peaks on the PHYLPRO profile (Fig. 2). We also suspect that there has been recombination within the human-specific clade because the topology was poorly supported and the MP retention index was only 0.39. The PHYLPRO profiles of the human-specific elements (analyzed on their own) fluctuated around 0.5, but with only an average of 1% nucleotide divergence between the sequences, it was not possible to localize and confirm specific recombination events.

This analysis demonstrates that there has been some recombination involving regions of a few hundred base pairs (probably as the result of gene conversion or switching between nonidentical templates during reverse transcription). However, it is clear that this recombination has had no significant effect on the overall phylogenetic signal for several reasons. PIST analysis found no evidence of recombination ($P = 0.154$), and the low informative sites index value of 0.036 indicates only a very low level of recombination (25). Furthermore, the MP trees constructed from the five different 856-bp data sets contained the same basic backbone structure (unpublished results), as did trees from 500-bp regions either side of the breakpoint at position 3223 (Fig. 2).

Mechanism of Proliferation. We find two lines of evidence for continuous purifying selection, which rules out a significant role for complementation in *trans* during the evolution of the HERV-K(HML2) family. First, inheritance of stop codons, which would be expected under the complementation in *trans* mechanism, is

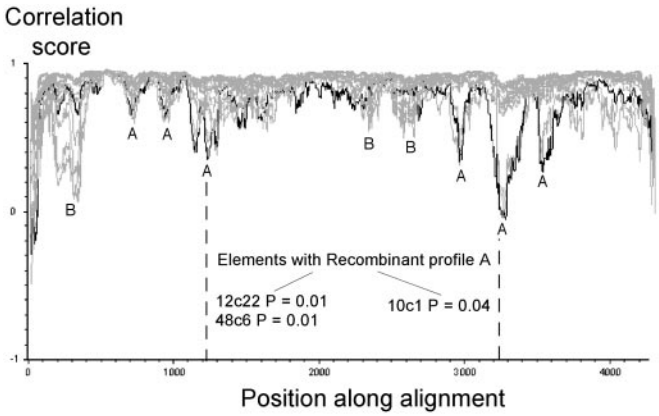

Fig. 2. Phylogenetic profile obtained by using PHYLPRO. The darkened line shows the profile for sequence 12c22 only. Negative peaks for groups A and B (members of which have almost identical profiles) are indicated. The vertical dashed lines extend from the break-point positions identified by LARD for 12c22 and the two other group A elements (HKY model, with κ and α estimated from neighbor joining trees for each triplet). *P* values are calculated from SEQ-GEN using 100 replicates from a Bayesian tree (HKY model: κ, α and base frequencies taken from the tree). The window size in PHYLPRO extends to include 100 variable positions either side of the sliding point. The ordinate shows the linear correlation score of uncorrected pairwise distances from the two windows.

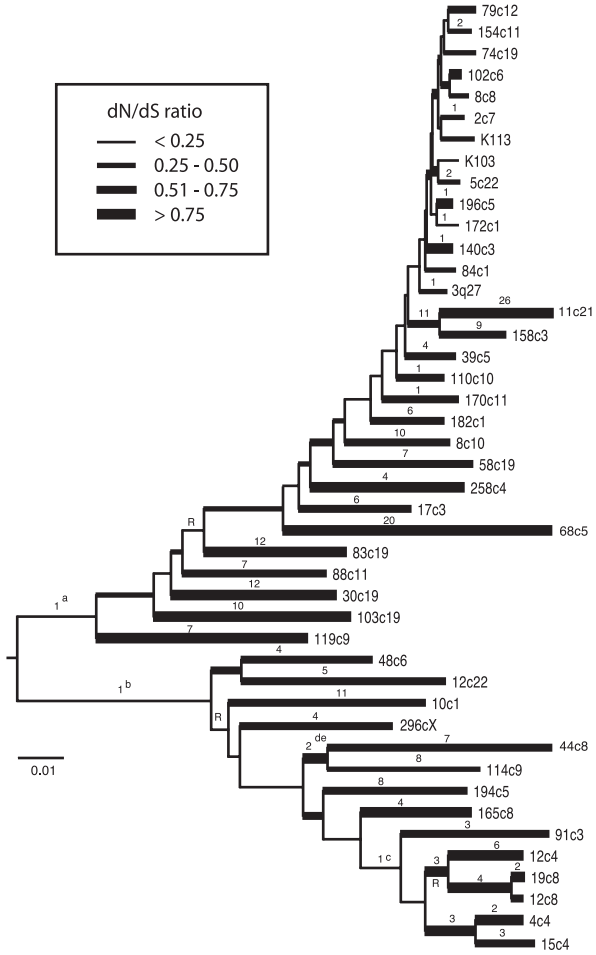

Fig. 3. Phylogeny of the internal region. The estimated number of stop codons acquired on each branch is shown (with *R* representing a reversal) and the d_N/d_S ratio is indicated by the line thickness. Superscript letters show the five potential acquisitions of stop codons on internal branches discussed in the text. Note that a short branch may represent insufficient changes for a reliable estimate to be made for that branch.

rare (Fig. 3). Excluding the segmentally duplicated clade, significantly fewer stop codons have been acquired on internal branches when compared to terminal branches, assuming an equal probability of acquisition per unit branch length (*G* test; $P < 0.001$). The major exception to this pattern is the branch joining the highly degenerate elements 11c21 and 158c3 (which have not been copied by segmental duplication). This branch accounts for 11 of the 16 inferred acquisitions of stop codons on internal branches and is our only definite example of complementation in *trans*. Although there are a further five stop codons present on internal branches (a–e in Fig. 3), we do not believe that any represent actual copying of defective elements: the summed likelihood difference caused by constraining each of them to be acquired on terminal branches is not significant $[2(l_1 - l_0) = 7.9, 5 \text{ df}, P = 0.16]$.

A second method by which purifying selection can be inferred is by calculating the ratio of nonsynonymous to synonymous changes (d_N/d_S ratio) on the different branches of the phylogeny (20). Mutations that lead to an amino acid substitution (nonsynonymous) are more likely to be deleterious to the protein's function than mutations that do not (synonymous), and hence, d_N/d_S ratios <1 indicate that purifying selection has been operating on the sequences since their divergence (values >1 indicate positive selection) (28). Conversely, if no selection has been operating on the sequences since their divergence, then d_N and d_S should be equal and the ratio should be 1, as is the case for most processed pseudogenes in the human genome (29).

We found that the d_N/d_S ratio in the HERV-K(HML2) family overall is low (0.52, Table 1) and, importantly, is generally lower on internal branches than on terminal ones (0.21 and 0.67 respectively, Table 1 and Fig. 3). This difference between internal and terminal branches was found to be highly significant with both the Wilcoxon rank sum test ($P < 0.001$) and likelihood ratio test (comparing the likelihoods of the one-ratio model and the two-ratio model, where the internal and terminal branches can have different values; Table 1). A striking exception to this pattern is the clade that has arisen as a result of segmental duplication of the human genome, where (in addition to having acquired 10 stop codons) the internal branches have the predicted d_N/d_S ratio of ≈ 1 (Table 3, which is published as supporting information on the PNAS web site). Similarly, the d_N/d_S ratio is high on the internal branch containing 11 stop codons. Thus, the generally low d_N/d_S ratio on the internal branches of our phylogeny supports our inference from the rarity of inherited stop codons that complementation in *trans* has been rare during the evolution of the HERV-K(HML2) family.

Retrotransposition in *cis* also appears to be rare because the *env* gene, which is only essential for infectious transmission between cells, also has few inherited stop codons (only e in Fig. 3), and a similarly reduced d_N/d_S ratio on internal compared to terminal branches of 0.26 and 0.79, respectively. These differences were also highly significant (Wilcoxon rank sum test, $P < 0.001$, and Table 1).

Interestingly, the observed level of purifying selection on the *env* gene on internal branches is not markedly dissimilar to that present within a clade of exogenous and infectious β retrovirses, including simian retrovirus types 1 and 2 and Mason–Pfizer monkey virus (Table 1). Furthermore, we find a similar pattern in the *env* gene within other HERV families (Table 2). Seven of the eight families have an *env* d_N/d_S ratio on internal branches that was significantly <1, and in several cases the ratio is comparable to that of the exogenous viruses (Table 1). Exceptions to this include a segmentally duplicated clade within the

EVOLUTION

Table 1. d_N/d_S ratios in the HERV-K(HML2) family and related exogenous β retroviruses

Family	Gene	One-ratio model (=1)	Two-ratio model (=2)			
		Whole tree	Internal branches*	Terminal branches	Likelihood difference (2 − 1)	*P* value†
HERV-K(HML2) (44 elements)	All	0.52	0.21	0.67	142	<0.001
	gag	0.53	0.28	0.68	26	<0.001
	pol	0.47	0.16	0.62	91	<0.001
	env	0.62	0.26	0.79	31	<0.001
Exogenous viruses (7 elements)	All	0.10	0.10	0.10	0.01	NS
	gag	0.09	0.09	0.10	0.05	NS
	pol	0.07	0.06	0.08	0.67	NS
	env	0.17	0.14	0.18	0.85	NS

*Excluding the segmentally duplicated clade (see Table 3).
†Taken from critical values on the χ^2 distribution. NS, not significant.

HERV-T family, and HERV-H elements that share inactivating deletions (30). In these two cases (where the elements are assumed to be evolving neutrally), the d_N/d_S ratios of internal branches are ≈1, as expected (Table 3 and Fig. 4, which is published as supporting information on the PNAS web site).

Consistent with previous results (26), the type I HERV-K(HML2) subgroup, defined by having the same 292-bp deletion spanning the *pol–env* boundary, is polyphyletic (Fig. 1). However, the deletion itself appears to have a single origin, because type I elements were monophyletic in a tree constructed from a data set derived from 400 bp either side of the deletion site. We therefore believe that recombination or gene conversion of relatively small fragments (which we have demonstrated) has acted to disperse the deletion among various members of the HERV-K(HML2) family, as previously suggested (26). The deletion in the type I elements alters the splicing pattern in *env* and might therefore be assumed to be inactivating. However, there is no significant difference in the *env* d_N/d_S ratio of the internal branches of phylogenies built from type I compared to type II elements [$2(l_1 - l_0) = 1.2$, 1 df, $P > 0.1$], suggesting that the *env* genes of the type I elements have been under purifying selection since the deletion event. Unfortunately, because of the lack of sequence divergence, we were unable to establish whether the 800-bp region spanning the deletion site is itself under selection. It is therefore possible (despite the disruption of *cORF* as a result of the deletion event; ref. 31) that the type I elements remain potentially functional. An alternative possibility is that the type I elements are nonfunctional and that (following its single origin), a relatively small region containing this deletion has proliferated by gene conversion.

Table 2. d_N/d_S ratios of the *env* gene on internal branches of trees from representative HERV families

Class	Family	No. of elements	d_N/d_S	*P* value*
I	HERV-E	26	0.22	<0.001
	HERV-H†	60	0.24	<0.001
	HERV-T‡	12	0.12	<0.001
	HERV-R	9	0.45	0.05–0.1
II	HERV-HML3	91	0.73	<0.05
	HERV-HML5	30	0.40	<0.001
	HERV-HML6	15	0.35	<0.001
III	HERV-S	34	0.21	<0.001

The significance of the difference in likelihood if this ratio is fixed at 1 is shown.
*Taken from critical values on the χ^2 distribution.
†Only for HERV-H elements that have an intact *env*.
‡Excluding elements copied by segmental duplication; this family is also termed HERV-S71.

Discussion and Conclusions

The paucity of inherited stop codons, and the low d_N/d_S ratios for all genes (including *env*) within the internal branches of the HERV-K(HML2) phylogeny, strongly indicate purifying selection, which in turn suggests that this family has increased in copy number predominantly by reinfection rather than by retrotransposition in *cis*, or complementation in *trans*. Proliferation via the latter two mechanisms would result in both the presence of numerous shared stop codons within the *env* gene and also in the neutral evolution of this gene (with a corresponding d_N/d_S ratio close to one), but this is not the case. Such reinfection may only involve movement from somatic to germ-line cells within the same individual, and does not necessarily require transfer between different individuals in the host population.

The apparent rarity of complementation in *trans* may seem surprising given our understanding of the mechanism of retroviral replication (32). One possible explanation is that often there is only a single element expressed in any particular cell, and so complementation in *trans* is not usually possible. The apparent rarity of retrotransposition in *cis* (as opposed to between-cell reinfection) is also surprising, and may suggest that gene expression is more rigidly controlled in germ-line cells than in the surrounding somatic cells.

Previously, Costas (26) demonstrated low d_N/d_S ratios in pairwise comparisons of genes from selected HERV-K(HML2) elements and suggested that there had been some transpositional activity since the human chimpanzee divergence some six million years ago. Here, however, in addition to showing that proliferation is largely caused by reinfection, our analyses allow us to make additional inferences about the evolution of the HERV-K(HML2) family. Continuous selection strongly suggests continuous functionality, and this implies that the HERV-K(HML2) lineage has retained replication-competent members since its origin. Furthermore, because selection is a property of a population that involves replication and loss of unfit elements, a pool of active elements must have been present throughout this period. Thus, the internal branches in our phylogeny do not represent the same element at different times, or single copying events. Rather, they represent samples from a changing pool of endogenous retroviruses (evolving primarily via viral rather than host DNA replication) that have been both replication competent and infectious throughout the evolution of the HERV-K(HML2) family. Most members of this pool would not have been fixed (i.e., they were only present within some individuals of the population) and there would have been a continuous turnover as elements were lost via natural selection and genetic

drift, and gained via new germ-line integrations. By contrast, the terminal branches in our phylogeny represent elements that have become both inactive and fixed, presumably by means of genetic drift. Thus, they account for almost all of the stop codon acquisitions and have a much higher d_N/d_S ratio. Even so, the terminal branches still have a d_N/d_S ratio of <1, possibly because many elements have been lost (via recombinational deletion or genetic drift), which will have merged internal branches onto terminal branches.

Although our phylogeny shows a rapid increase in the number of elements following the origin of humans, and changes shape from wholly pectinate (comb-shaped) to more balanced, this should not be taken as evidence for a human-specific burst of activity. We suggest that the change in tree topology has been caused simply by there having been less time for recent elements to be lost by recombinational deletion (leaving only solo LTRs). In the future, when most of the modern elements have undergone such events, the phylogeny of those remaining is likely to be sparse and pectinate, as the phylogeny of the older elements is now. We found evidence of this process occurring at the present time in the relatively young, human-specific element K103 (13). Although most humans carry the full-length virus, we found that the human genome project sequence and several sub-Saharan African individuals have only a solo LTR at the integration site (unpublished data).

The scenario of continuous purifying selection described above explains why the human genome still contains relatively intact HERV-K(HML2) elements, even after a 30-million-year-long association with this particular viral lineage. Previously, such elements were thought to result from the mutational or recombinational reactivation of elements that were otherwise continually decaying due to errors introduced during host replication (7, 33). Our examination of *env* d_N/d_S ratios in other HERV families suggests that the dominant role of reinfection and the persistence of a pool of active elements may be a general feature of HERV evolution. However, we note that one HERV family (HERV-L) does not encode an *env* gene, and that in another (HERV-H), >90% of the elements share large deletions in *pol* and *env* (30). Thus, in these cases (which are the two largest families in terms of copy number), proliferation has probably occurred via alternative mechanism(s) to reinfection.

Until now, our understanding of the evolution of interspersed repeats such as HERVs has been influenced heavily by phylogenetic tree shape (34), and the typically unbalanced phylogenies have been thought to reflect one or a few active elements (called masters) giving rise to many other copies that do not copy themselves (35). In such models, applied also to the HERV-K(HML2) family (26, 36), the internal nodes represent the same master element (at the same integration site) at different times. In these scenarios, the master elements can remain functional for long periods of time, but they would still be expected to accumulate synonymous and nonsynonymous substitutions at the same rate. Hence, the d_N/d_S ratios on the internal branches of the HERV-K(HML2) phylogeny would not be significantly different from 1. In contrast, our results show that the internal branches are under strong purifying selection, and internal nodes represent different elements that are survivors from a pool of unfixed, active endogenous retroviruses. Thus, although there are no HERV-K(HML2) elements in the completed human genome sequence with a full coding capacity (which has led to the conclusion that this family is unlikely to be capable of causing disease; refs. 7 and 33), our data indicate that, as suggested by previous authors (14, 26), a proportion of the human population may still harbor active and infectious members of the HERV-K(HML2) family.

We thank James Cook and Donald Quicke for comments on the manuscript. This work forms part of a project funded by the Wellcome Trust. A.K. and V.P. were in receipt of Natural Environment Research Council studentships, and J.P. was supported by Centre for Integrated Genomics Grant LN00A079.

1. Boeke, J. D. & Stoye, J. P. (1997) in *Retroviruses*, eds. Coffin, J. M., Hughes, S. H. & Varmus, H. E. (Cold Spring Harbor Lab. Press, Plainview, NY), pp. 343–435.
2. International Human Genome Sequencing Consortium (2001) *Nature* **409,** 860–921.
3. Pačes, J., Pavlíček, A. & Pačes, V. (2002) *Nucleic Acids Res.* **30,** 205–206.
4. Tristem, M. (2000) *J. Virol.* **74,** 3715–3730.
5. Bénit, L., Dessen, P. & Heidmann, T. (2001) *J. Virol.* **75,** 11709–11719.
6. Katzourakis, A. & Tristem, M. (2004) in *Retroviruses and Primate Genome Evolution*, ed. Sverdlov, E. D. (Landes Bioscience, Georgetown, TX).
7. Stoye, J. P. (2001) *Curr. Biol.* **11,** R914–R916.
8. Wei, W., Gilbert, N., Ooi, S. L., Lawler, J. F., Ostertag, E. M., Kazazian, H. H., Boeke, J. D. & Moran, J. V. (2001) *Mol. Cell. Biol.* **21,** 1429–1439.
9. Song, S. U., Kurkulos, M., Boeke, J. D. & Corces, V. G. (1997) *Development (Cambridge, U.K.)* **124,** 2789–2798.
10. Löwer, R., Löwer, J. & Kurth, R. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 5177–5184.
11. Andersson, M. L., Lindeskog, M., Medstrand, P., Westley, B., May, F. & Blomberg, J. (1999) *J. Gen. Virol.* **80,** 255–260.
12. Medstrand, P. & Mager, D. L. (1998) *J. Virol.* **72,** 9782–9787.
13. Barbulescu, M., Turner, G., Seaman, M. I., Deinard, A. S., Kidd, K. K. & Lenz, J. (1999) *Curr. Biol.* **9,** 861–868.
14. Turner, G., Barbulescu, M., Su, M., Jensen-Seaman, M. I., Kidd, K. K. & Lenz, J. (2001) *Curr. Biol.* **11,** 1531–1535.
15. Hughes, J. F. & Coffin, J. M. (2001) *Nat. Genet.* **29,** 487–489.
16. Ronquist, F. & Huelsenbeck, J. P. (2003) *Bioinformatics*, **19,** 1572–1574.
17. Swofford, D. L. (1998) PAUP*: *Phylogenetic Analysis Using Parsimony (*and Other Methods)* (Sinauer, Sunderland, MA), Version 4.
18. Pagel, M. (1999) *Nature* **401,** 877–884.
19. Yang, Z. H. (1997) *Comput. Appl. Biosci.* **13,** 555–556.
20. Yang, Z. (1998) *Mol. Biol. Evol.* **15,** 568–573.
21. Rice, P. Longden, I. & Bleasby, A. (2000) *Trends Genet.* **16,** 276–277.
22. Weiller, G. F. (1998) *Mol. Biol. Evol.* **15,** 326–335.
23. Holmes, E. C., Worobey, M. & Rambaut, A. (1999) *Mol. Biol. Evol.* **16,** 405–409.
24. Rambaut, A. & Grassly, N. C. (1997) *Comput. Appl. Biosci.* **13,** 235–238.
25. Worobey, M. (2001) *Mol. Biol. Evol.* **18,** 1425–1434.
26. Costas, J. (2001) *J. Mol. Evol.* **53,** 237–243.
27. Johnson, W. E. & Coffin, J. M. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 10254–10260.
28. Li, W.-H. (1997) *Molecular Evolution* (Sinauer, Sunderland, MA).
29. Bustamante, C. D., Nielsen, R. & Hartl, D. L. (2002) *Mol. Biol. Evol.* **19,** 110–117.
30. Mager, D. L. & Freeman, J. D. (1995) *Virology* **213,** 395–404.
31. Löwer, R, Tönjes, R. R., Korbmacher, C., Kurth, R. & Löwer, J. (1995) *J. Virol.* **69,** 141–149.
32. Swanstrom, R. & Wills, J. W. (1997) in *Retroviruses*, eds. Coffin, J. M., Hughes, S. H. & Varmus, H. E. (Cold Spring Harbor Lab. Press, Plainview, NY), pp. 263–334.
33. Gifford, R. & Tristem, M. (2003) *Virus Genes* **26,** 291–315.
34. Clough, J. E., Foster, J. A., Barnett, M. & Wichman, H. A. (1996) *J. Mol. Evol.* **42,** 52–58.
35. Deininger, P. L., Batzer, M. A., Hutchison, C. A. & Edgell, M. H. (1992) *Trends Genet.* **8,** 307–311.
36. Buzdin, A., Ustyugova, S., Khodosevich, K., Mamedov, I., Lebedev, Y. Hunsmann, G. & Sverdlov, E. (2003) *Genomics*, **81,** 149–156.