doi:10.1093/nar/gkm697

The *Gypsy* Database (GyDB) of mobile genetic elements

C. Lloréns[1,2,*], R. Futami[1], D. Bezemer[3] and A. Moya[2,4]

[1]Biotech Vana, Valencia, [2]Institut Cavanilles de Biodiversitat i Biología Evolutiva Universitat de València, Spain, [3]HIV Monitoring Foundation, Amsterdam, The Netherlands and [4]CIBER de Epidemiología y Salud Pública (CIBERESP), Spain

Received June 29, 2007; Revised August 19, 2007; Accepted August 22, 2007

ABSTRACT

In this article, we introduce the Gypsy Database (GyDB) of mobile genetic elements, an in-progress database devoted to the non-redundant analysis and evolutionary-based classification of mobile genetic elements. In this first version, we contemplate eukaryotic *Ty3/Gypsy* and *Retroviridae* long terminal repeats (LTR) retroelements. Phylogenetic analyses based on the gag-pro-pol internal region commonly presented by these two groups strongly support a certain number of previously described *Ty3/Gypsy* lineages originally reported from reverse-transcriptase (RT) analyses. Vertebrate retroviruses (*Retroviridae*) are also constituted in several monophyletic groups consistent with genera proposed by the ICTV nomenclature, as well as with the current tendency to classify both endogenous and exogenous retroviruses by three major classes (I, II and III). Our inference indicates that all protein domains codified by the gag-pro-pol internal region of these two groups agree in a collective presentation of a particular evolutionary history, which may be used as a main criterion to differentiate their molecular diversity in a comprehensive collection of phylogenies and non-redundant molecular profiles useful in the identification of new *Ty3/Gypsy* and *Retroviridae* species. The GyDB project is available at http://gydb.uv.es.

INTRODUCTION

Since the existence of mobile DNA was first suggested by McClintock (1), mobile genetic elements have been an important object of study in multiple areas of biological research (2). Mobile genetic elements are self-contained genomic units capable of proliferating within their host genomes. Nearly all fit into three major functional categories: Class I are all reverse-transcriptase (RT) dependent retroelements (3) that mediate their transposition life cycle through an RNA–DNA reverse transcription process; Class II are DNA-based transposons that move directly from one position to another in host genomes (1,4,5) and Class III are the miniature inverted-repeats transposable elements (MITEs) (6,7). With continuous efforts in sequencing and annotation, the field of genomics has been dramatically expanded in the attempt to understand the gene organization of genomes, as well as the bioinformatic and empirical characterization of open reading frames (ORFs). Most of these efforts have revealed mobile genetic elements to be more widely distributed in the genomes of eukaryotes than previously thought; it is thus, commonly accepted that they may have played an important role in the evolution of life and the origin of eukaryotic complexity (8). With the aim of furthering knowledge in this field, we have built the GyDB, a research project in which we analyze and classify non-redundant mobile genetic elements based on their evolutionary profiles. Due to their impressive molecular diversity, the GyDB is a long-term project that has been arranged in a database in continuous progress and must be achieved in stages. In this article, we introduce the database and its background focusing on *Ty3/Gypsy* and *Retroviridae* long terminal repeats (LTR) retroelements (LTR retrotransposons and retroviruses). The database also focuses on certain non-viral protein families related to these two groups.

Ty3/Gypsy and *Retroviridae* related websites

The *Retroviridae* are viral particles that reverse-transcribe their RNA genome into a double-stranded DNA copy inserted in the infected host cell genome. Their diploid RNA genome is enveloped within a protein capsid (CA) by a membrane fragment of the host cell in which envelope (env) antigens are embedded. Vertebrate retroviruses initially received attention with the description of the oncogenic human T-cell leukemia virus (HTLV-I), the first retrovirus found to be pathogenic in humans (9,10), and, later, with the discovery of the human immunodeficiency virus type 1 (HIV-1), the agent

[*] To whom correspondence should be addressed. Tel: +34 963 553 182; Fax: +34 963 561 641; Email: carlos.llorens@uv.es

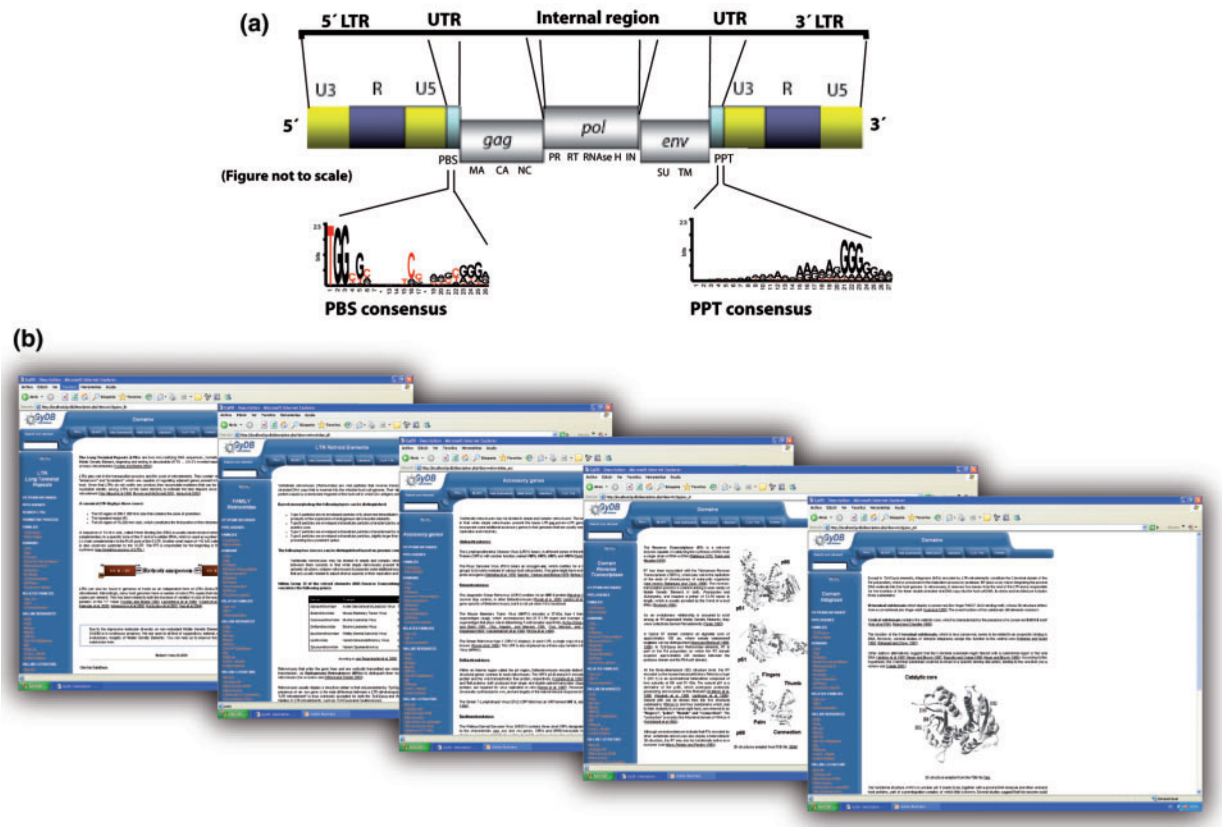

Figure 1. (**a**) Genomic structure of a basal retrovirus, and logos to graphically represent the consensus for both the PBS and the PPT motifs. (**b**) Screenshot of the GyDB websites specific to families and protein domains.

responsible for acquired immune deficiency syndrome (AIDS) (11–13). There are at present 15–25 million people worldwide infected with the HTLV-1 (14), and nearly 40 million with the HIV (15). *Ty3/Gypsy* LTR retroelements are mobile genetic elements that mediate their transposition cycle through an RNA–DNA reverse transcription process, they were originally described as retrotransposable sequences present in the genomes of yeasts and flies (16–18), and are similar to vertebrate retroviruses in LTR-gag-pol-LTR genomic structure and sequence. The main difference between a retrovirus and a canonical LTR retrotransposon is thus that retroviruses have an additional ORF encoding for an env polyprotein necessary for transferring retroviruses from cell to cell. However, currently it is well-known that *env*-like genes are not exclusive of vertebrate retroviruses (19), and since many studies converged in disclosing that certain *Ty3/Gypsy* and other LTR retroelement lineages are well functional as well as potential retroviruses (20–26) the possibility that any LTR retrotransposon could become a potential retrovirus when acquiring an *env* gene is a fascinating object of research. Figure 1a summarizes the structure of a *Ty3/Gypsy* or *Retroviridae* simple retrovirus, which is characterized by an internal region flanked by two normally homologous non-coding DNA sequences named LTRs. The internal region contains three ORFs arranged in the following order (27); first, a *gag* gene coding for a gag precursor containing the matrix (MA), CA and nucleocapsid (NC) domains; second, a *pol* gene coding for a pol polyprotein, which usually contains the protease (PR), RT, ribonuclease H (RNAse H) and integrase (INT) domains and third, the *env* gene coding for an env glycoprotein containing the outer surface (SU) membrane protein and the transmembrane (TM) protein. Both *Ty3/Gypsy* and *Retroviridae* families, species, as well as LTRs and protein domains, have within the GyDB a website that provides a brief discussion, structural representations and bibliographic references, as shown in Figure 1b.

Phylogenetic analyses: clades and genera

The first version of the GyDB focuses on the exhaustive analysis of 120 non-redundant *Ty3/Gypsy* and *Retroviridae* full-length genomes collected at the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/). The most conserved part (core) of each protein domain was aligned using CLUSTALX (28) and refined with GENEDOC editor (http://www.psc.edu/biomed/genedoc). Although the *Retroviridae* display identical gag-pro-pol-env structure as *Ty3/Gypsy* retroviruses (29) not all *Ty3/Gypsy* LTR retroelements are retroviruses, and it is well supported that the different lineages of retroviruses described in invertebrates probably acquired their *env* genes by independent gene recruitment events (see Ref. (29) and references therein). Consequently, the most valuable relationships between *Ty3/Gypsy* and *Retroviridae* LTR retroelements should be sought in the internal region that codifies for the gag and pol polyproteins. The criteria for LTR retroelement classification at the GyDB are thus based on the clusters reported by a majority-rule consensus (MRC) tree inferred based on a concatenated gag-pro-pol multiple alignment containing the most conserved part of the CA, NC, PR, RT, RNAseH and INT domains. Nevertheless, we have also inferred and provide online, independent phylogenies based on the gag polyprotein, the pol polyprotein and all pol protein domains, and the env polyprotein. The gag-pro-pol alignment has therefore two components, the gag polyprotein and the pol polyprotein. Regarding the gag polyprotein we consider only the CA–NC region because MA is absent in many *Ty3*/*Gypsy* sequences and in others cannot be exhaustively aligned due to extreme divergence. Concerning the pol polyprotein, we consider the PR-RT-RNAseH-INT region from the catalytic DTG PR motif (30) to the GPY/F INT module (31). The PR domain is taken as another pol component as it has a low but similar phylogenetic signal than other pol protein domains (see PR MRC tree in the 'Section Phylogenies', at GyDB). As shown in Figure 2, gag-pro-pol tree agrees and improves all clades and genera heretofore inferred based on the RT, RNAseH or INT pol-like domains (22–24,26,31–45). This indicates that despite the different rates of evolution (not considered by parsimony method) all protein domain encoded by the gag-pro-pol internal region (except MA) have a similar phylogenetic signal that may be used as a main criterion to phylogenetically classifying and profiling the currently known *Ty3*/*Gypsy* and *Retroviridae* diversity. In an attempt to identify the most satisfactory method of phylogenetic inference, we tested the distance-based neighbour-joining (NJ) method (46) and the minimum-change-based Parsimony method (47,48) using Phylip 3.6 (http://evolution.gs.washington.edu/phylip.html) to infer MRC trees (49). The two methods reported identical clusters of operative taxonomical units (OTUs) (see Llorens and Moya, the Three Kings Hypothesis, manuscript in preparation). This has allowed us to taxonomically and realistically define the monophyletic clusters of protein families, independently of which method would be used. However, the parsimony method was revealed to be much more consistent with comparative analyses than NJ-method when inferring phylogenies based on non-conserved protein domains such as the gag polyprotein and the protease domain. Although these two proteins are extremely divergent (less than 20% of overall identity), all sequences belonging to a particular lineage have an amino acid architecture in common that is similar but divergent from that displayed in other lineages. The point is that when inferring phylogenies involving these two proteins, parsimony method always anticipated in our analyses a MRC tree more consistent with comparative analyses than NJ, and also supported the overall clustering with better statistical values. We have thus chosen Parsimony MRC trees as principal phylogenetic reference, at GyDB. Phylogeny websites are presented through an HTML file where clicking on the name of any retroelement, will access a link to a descriptive file that in turn links to the NCBI Genbank accession of the requested element, as well as a short discussion, taxonomy information, genomic structure and a bibliography concerning the element described. If the selected element has no file, the link takes the user directly to the sequence's Genbank accession at the NCBI.

Retroviridae accessory genes

Vertebrate retroviruses may be divided into simple and complex retroviruses. The main distinction is that while simple retroviruses present the basal LTR-gag-pol-env-LTR genomic structure, complex retroviruses incorporate in their genomes additional accessory genes usually needed to adjust diverse aspects of their replication and infectivity. Table 1 summarizes a list of the accessory genes that may be characteristic of a genus, characteristic of a clade within a genus, and in certain cases exclusive to a unique retrovirus; we provide a brief discussion of each accessory gene and bibliographic references within the accessory genes website, at GyDB (http://gydb.uv.es/gydb/description.php?desc=retroviridae_acc). Accessory genes phylogenies are available online together with the other phylogenetic reconstructions in the section 'Phylogenies' of the database.

Related families of non-viral proteins

It is well known that several protein domains encoded by retroelements in general are related to certain families of non-viral proteins present in the genomes of eukaryotes and prokaryotes. It is thus commonly accepted that these kinds of proteins have an ancient relationship with retroelements. The origin of mobile genetic elements, as well as their role in the evolution of eukaryotic complexity, is thus a fascinating subject of discussion and controversy. We are particularly interested in this topic and have considered in this first version or our database the following three non-viral protein families related to LTR retroelements: chromodomains (50), GIN-1 integrases (51) and clan AA of aspartic peptidases (52). Each of these has its own website and phylogeny within the GyDB.

BLAST and HMM servers

One of the most important goals of our project is to provide a set of competent services to facilitate the identification and taxonomical classification of new retroelement species. In an attempt to support further sequence–sequence identification, we have implemented a BLAST search (53) that allows the typical comparisons to the following databases: LTR, GENOME and CORES. These databases respectively contain the LTR nucleotide sequences, the complete element genome and the core of each detectable protein domain encoded by the LTR retroelements we currently classify. Results are reported in the conventional BLAST output. However, similarities detected by an unknown query are identified by the name of the element to which the detected sequence belongs, and provide a link to the sequence's Genbank accession. The GyDB BLAST databases are non-redundant, and specific. This facilitates the analysis of pairwise similarities among both closely and distantly

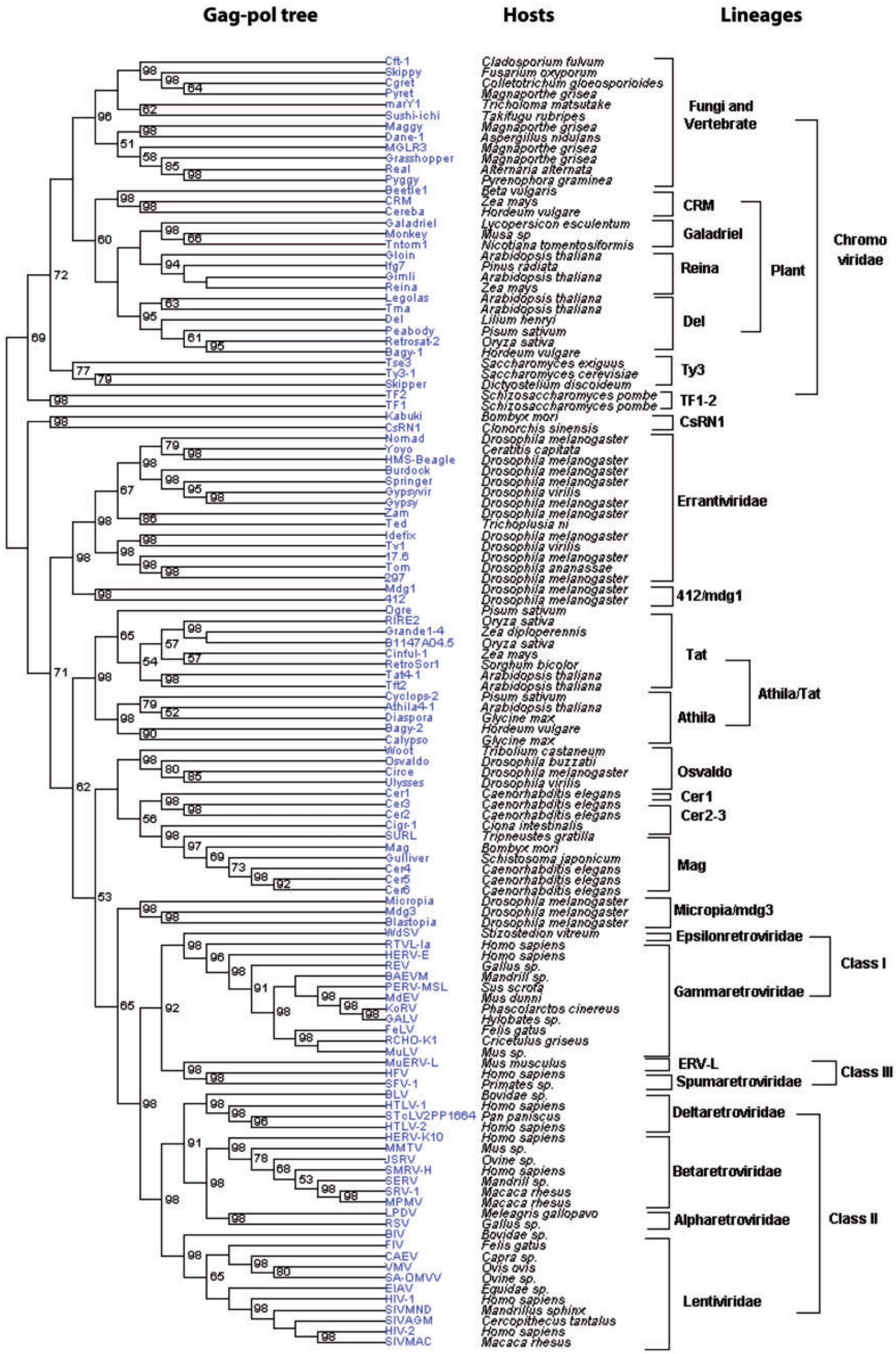
Gag-pol tree
Hosts
Lineages
Cft-1 Cladosporium fulvum
Skippy Fusarium oxyporum
Cgret Colletotrichum gloeosporioides
Pyret Magnaporthe grisea
marY1 Tricholoma matsutake
Sushi-ichi Takifugu rubripes
Maggy Magnaporthe grisea
Dane-1 Aspergillus nidulans
MGLR3 Magnaporthe grisea
Grasshopper Magnaporthe grisea
Real Alternaria alternata
Pyggy Pyrenophora graminea
Beetle1 Beta vulgaris
CRM Zea mays
Cereba Hordeum vulgare
Galadriel Lycopersicon esculentum
Monkey Musa sp
Tntom1 Nicotiana tomentosiformis
Gloin Arabidopsis thaliana
Ifg7 Pinus radiata
Gimli Arabidopsis thaliana
Reina Zea mays
Legolas Arabidopsis thaliana
Tma Arabidopsis thaliana
Del Lilium henryi
Peabody Pisum sativum
Retrosat-2 Oryza sativa
Bagy-1 Hordeum vulgare
Tse3 Saccharomyces exiguus
Ty3-1 Saccharomyces cerevisiae
Skipper Dictyostelium discoideum
TF2 Schizosaccharomyces pombe
TF1 Schizosaccharomyces pombe
Kabuki Bombyx mori
CsRN1 Clonorchis sinensis
Nomad Drosophila melanogaster
Yoyo Ceratitis capitata
HMS-Beagle Drosophila melanogaster
Burdock Drosophila melanogaster
Springer Drosophila melanogaster
Gypsyvir Drosophila virilis
Gypsy Drosophila melanogaster
Zam Drosophila melanogaster
Ted Trichoplusia ni
Idefix Drosophila melanogaster
Tv1 Drosophila virilis
17.6 Drosophila melanogaster
Tom Drosophila ananassae
297 Drosophila melanogaster
Mdg1 Drosophila melanogaster
412 Drosophila melanogaster
Ogre Pisum sativum
RIRE2 Oryza sativa
Grande1-4 Zea diploperennis
B1147A04.5 Oryza sativa
Cinful-1 Zea mays
RetroSor1 Sorghum bicolor
Tat4-1 Arabidopsis thaliana
Tft2 Arabidopsis thaliana
Cyclops-2 Pisum sativum
Athila4-1 Arabidopsis thaliana
Diaspora Glycine max
Bagy-2 Hordeum vulgare
Calypso Glycine max
Woot Tribolium castaneum
Osvaldo Drosophila buzzatii
Circe Drosophila melanogaster
Ulysses Drosophila virilis
Cer1 Caenorhabditis elegans
Cer3 Caenorhabditis elegans
Cer2 Caenorhabditis elegans
Cigr-1 Ciona intestinalis
SURL Tripneustes gratilla
Mag Bombyx mori
Gulliver Schistosoma japonicum
Cer4 Caenorhabditis elegans
Cer5 Caenorhabditis elegans
Cer6 Caenorhabditis elegans
Micropia Drosophila melanogaster
Mdg3 Drosophila melanogaster
Blastopia Drosophila melanogaster
WdSV Stizostedion vitreum
RTVL-Ia Homo sapiens
HERV-E Homo sapiens
REV Gallus sp.
BAEVM Mandrill sp.
PERV-MSL Sus scrofa
MdEV Mus dunni
KoRV Phascolarctos cinereus
GALV Hylobates sp.
FeLV Felis gatus
RCHO-K1 Cricetulus griseus
MuLV Mus sp.
MuERV-L Mus musculus
HFV Homo sapiens
SFV-1 Primates sp.
BLV Bovidae sp.
HTLV-1 Homo sapiens
STcLV2PP1664 Pan paniscus
HTLV-2 Homo sapiens
HERV-K10 Homo sapiens
MMTV Mus sp.
JSRV Ovine sp.
SMRV-H Homo sapiens
SERV Mandrill sp.
SRV-1 Macaca rhesus
MPMV Macaca rhesus
LPDV Meleagris gallopavo
RSV Gallus sp.
BIV Bovidae sp.
FIV Felis gatus
CAEV Capra sp.
VMV Ovis ovis
SA-OMVV Ovine sp.
EIAV Equidae sp.
HIV-1 Homo sapiens
SIVMND Mandrillus sphinx
SIVAGM Cercopithecus tantalus
HIV-2 Homo sapiens
SIVMAC Macaca rhesus
Fungi and Vertebrate
CRM
Galadriel
Reina
Del
Plant
Chromo viridae
Ty3
TF1-2
CsRN1
Errantiviridae
412/mdg1
Tat
Athila/Tat
Athila
Osvaldo
Cer1
Cer2-3
Mag
Micropia/mdg3
Epsilonretroviridae
Gammaretroviridae
Class I
ERV-L
Spumaretroviridae
Class III
Retro viridae
Deltaretroviridae
Betaretroviridae
Alpharetroviridae
Class II
Lentiviridae

Figure 2. MRC tree inferred for *Ty3*/*Gypsy* and *Retroviridae* LTR retroelements using the parsimony method and based on a concatenated gag-pro-pol multiple alignment. Host organisms and monophyletic clusters are detailed at left. MRC trees usually consist of all groups that occur more than 50% of the time, we take consensus values higher than 55 as an equivalent-bootstrapping reference.

Table 1. The Gypsy database. Accessory genes and complex retroviruses

Gene	Lineage	Specific of
orf1	*Alpharetroviridae*	Lymphoproliferative disease virus (LPDV)
orf2	*Alpharetroviridae*	Lymphoproliferative disease virus (LPDV)
orf3	*Alpharetroviridae*	Lymphoproliferative disease virus (LPDV)
orf4	*Alpharetroviridae*	Lymphoproliferative disease virus (LPDV)
src	*Alpharetroviridae*	Rous sarcoma virus (RSV)
bel1	*Spumaretroviridae*	Common for all spumaretroviruses
bel2	*Spumaretroviridae*	Common for all spumaretroviruses
bel3	*Spumaretroviridae*	Common for all spumaretroviruses
orfX	*Betaretroviridae*	Common for all betaretroviruses
sag	*Betaretroviridae*	Mouse mammary tumor virus (MMTV)
sorf	*Betaretroviridae*	Simian retrovirus type 1 (SRV-1)
rex	*Deltaretroviridae*	Common for all deltaretroviruses
rof	*Deltaretroviridae*	Common for all deltaretroviruses
tax	*Deltaretroviridae*	Common for all deltaretroviruses
tof	*Deltaretroviridae*	Common for all deltaretroviruses
orfV	*Deltaretroviridae*	Simian T-lymphotropic virus (STcLV2PP1664)
orfA	*Epsilonretroviridae*	Walleye dermal sarcoma virus (WDSV)
orfB	*Epsilonretroviridae*	Walleye dermal sarcoma virus (WDSV)
orfC	*Epsilonretroviridae*	Walleye dermal sarcoma virus (WDSV)
rev	*Lentiviridae*	Common for all lentiviruses
tat	*Lentiviridae*	Common for all lentiviruses
vif/sor/orfQ	*Lentiviridae*	All lentiviruses except EIAV
vpr	*Lentiviridae*	Primate lentiviruses except HIV-2 and certain relatives
vpx	*Lentiviridae*	Primate lentiviruses
nef	*Lentiviridae*	Primate lentiviruses
vpu	*Lentiviridae*	Human immunodeficiency viruses type-1 (HIV-1)
tmx	*Lentiviridae*	Bovine immunodeficiency virus (BIV)
vpw	*Lentiviridae*	Bovine immunodeficiency virus (BIV)
vpy	*Lentiviridae*	Bovine immunodeficiency virus (BIV)
orfs	*Lentiviridae*	Equine infectious Anemia virus (EIAV)
orfA	*Lentiviridae*	Feline immunodeficiency virus (FIV)
orfW	*Lentiviridae*	Visna viruses

related sequences with the same known function. On the other hand, Hidden Markov Model (HMM) profiles are statistical models that capture position-specific information on the degree of conservation in the DNA or protein domain architecture of an alignment and model the primary structure consensus of a family of protein or DNA sequences. Taking this into account we have also constructed, using HMMER Version 2.3.2 (54), a collection of HMM profiles considering for each protein domain a certain number of local multiple alignments extrapolated from the monophyletic clusters reported by the gag-pol-tree summarized in Figure 2. Our HMM profiles are part of the GyDB collection, which consists of a set of non-redundant multiple alignments, HMM profiles and MRC sequences, available to Biotech Vana registered users only (Biotech Vana Bioinformatics, in preparation). However, we implement a publicly available HMM server that, via HMMER, permits a user to search the entire HMM profile database with an unknown query or to search the CORES database using an HMM profile as a query. Outputs are generated in the usual style of HMMER, and allow users to easily identify the clade and/or genus to which a protein query taxonomically belongs.

Literature server

By way of this server users can access a database with citations specific to *Ty3/Gypsy* and *Retroviridae* LTR retroelements. The typical filters of year, journal, author and title may be applied in searches. Each displayed citation links to the PubMed Central digital archive at NCBI.

Database arrangement and navigation

The GyDB has been installed on a MySQL server. The server PHP language has been used to design the Web interface and service scripts that realize requests to the MySQL database, offering users a simple interaction and navigation facilitated by specially tailored search engines and an intuitively comprehensible menu. The whole system is implemented in a server based in a Linux environment and a Web Apache server. The navigation within the GyDB is notably intuitive. As shown in Figure 3, its foundation is a trio of Web browsers: element browser, menu and upper browser. The element browser is located to the left of the upper browser; it is a shortcut to accessing LTR retroelement files. Upon the introduction of a requested element's acronym, the element browser takes the user directly to an element file. The menu browser directs users to all GyDB websites. The upper browser provides access to the BLAST server, to a data submission form, to the HMM server, to the literature database and to a descriptive map on which Figure 3 is based.

Empirical example

In an attempt to provide an empirical example of the possibilities of our database, in this section we analyze

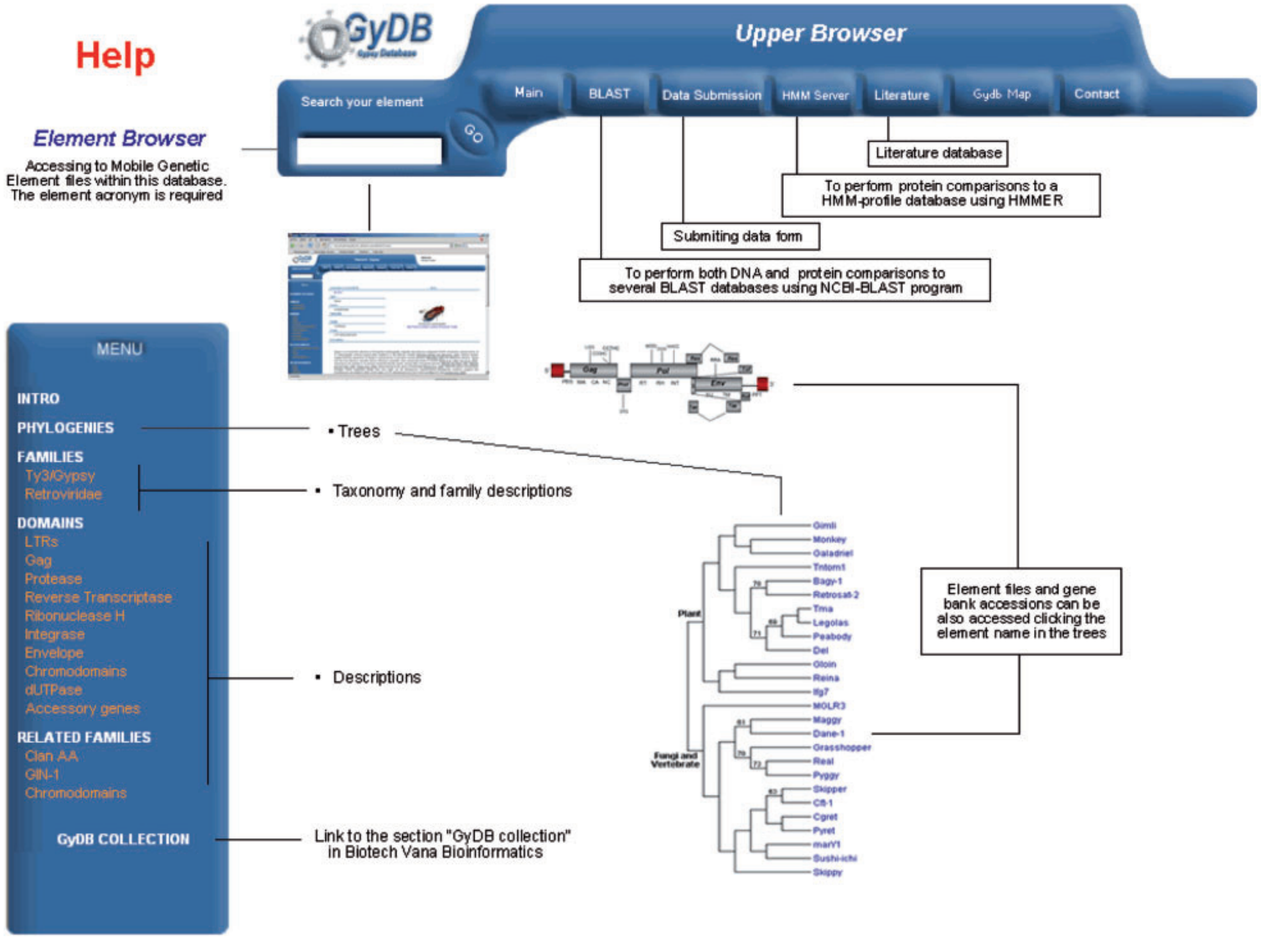

Figure 3. Database arrangement and navigation.

the recently described *Python molurus* retrovirus (PyERV), an endogenous retrovirus whose classification is unclear (55). According to the authors of this study, PyERV is a possible true recombinant related to B- and D-type retroviruses. From both viral taxonomy and morphology, it is now known that betaretroviruses may be divided into B- and D-type retroviruses (40). Also, it should be noted that although B- and D-type betaretroviruses are closely similar in the entire gag-pro-pol internal region, they differ in the env region. In this regard, it is well known that primates' D-type betaretroviruses present a common surface receptor also utilized by baboon and cat endogenous C-type gammaretroviruses (56,57). This evidence seems to be related to the high similarity displayed between env polyproteins encoded by gammaretroviruses and D-type betaretroviruses, where it is usually assumed that D-type betaretroviruses might be recombinant hybrids between C-type gammaretroviruses and primates' B-type betaretroviruses (40,58). With this, our profile database provides two independent HMM profiles independently describing the env polyproteins of B- and D-type betaretroviruses. Regarding PyERV, this retrovirus contains intact ORFs for the *gag*, *pro*, *pol* and *env* genes characteristic of retroviruses, and also an additional ORF of unknown function. Several comparisons were established against the HMM server using all protein domains encoded by PyERV as query examples (Genbank accession AF500296). Except in the case of the env polyprotein, where PyERV is slightly closer to gammaretroviruses than to D-type betaretroviruses (Table 2), all gag-pro-pol comparisons revealed that PyERV is clearly similar to betaretroviruses in general (Table 2). On the other hand, PyERV encodes for a dUTPase (DUT) domain, which is characteristic of betaretroviruses, non-primate lentiviruses and ERV-L elements (59). However, as it is also observed in betaretroviruses, PyERV-DUT is found in frame and N-terminal to the PR domain, while lentiviruses and ERV-L elements present this gene between or downstream to the RNAseH and INT domains. Analyses did not detect similarity between the unknown ORF described by the authors of PyERV study. However, immediately downstream to the same frame, PyERV codify for an amino acid stretch significantly similar to the putative ORF-X protein of betaretroviruses (Figure 4a). This is probably a frameshifting of the uncharacterized ORF described in PyERV by Huder *et al.* (55). ORF-X was originally described in the Jaagsiekte Sheep Retrovirus (JSRV) and other endogenous sheep betaretroviruses as a putative accessory gene that codifies for a protein similar

Table 2. Hits for protein family classification of the env polyprotein of PyERV

Alpha retroviridae		B-type betaretroviridae		D-type betaretroviridae		Gamma retroviridae		Delta retroviridae		Lenti viridae	
Score	*E*-value	Score	*E*-value	Score	*E*-value	Score	*E*-value	Score	*E*-value	Score	*E*-value
−112.4	0.0044	−127.8	0.0033	120.9	6.4*E*−35	482.3	1*E*−143	53.8	4.8*E*−18	8.9	0.00018

Table 3. Hits for protein family classification of the gag-pro-pol internal region of PyERV

Domain	Alpha retroviridae		Beta retroviridae		Gamma retroviridae		Delta retroviridae		Lenti viridae	
Query	Score	*E*-value	Score	*E*-value	Score	*E*-value	Score	*E*-value	Score	*E*-value
GAG	10.5	1.4*E*−05	116.1	9.9*E*−37	−0.9	0.061	27.9	2.3*E*−09	73.5	1.8*E*−24
DUT	no	no	−83.2	0.027	no	no	no	no	102.2	0.41
PR	16.6	1.4*E*−06	40.9	7.1*E*−11	1.8	0.83	5.5	0.0077	6.7	0.0077
RT	304.7	2.4*E*−90	393.3	5.4*E*−117	6.4	1.4*E*−20	144.6	3.9*E*−50	137.9	4.1*E*−40
RNAseH	49.2	5.3*E*−15	99.3	4.3*E*−30	1.3	0.049	16.5	6*E*−06	0.7	0.17
INT	160.9	6.2*E*−54	266.8	6.1*E*−79	5.2	0.0054	115.8	1.7*E*−33	79.6	6.2*E*−25
ORF X	no	no	22.6	3.8*E*−07	no	no	no	no	no	no

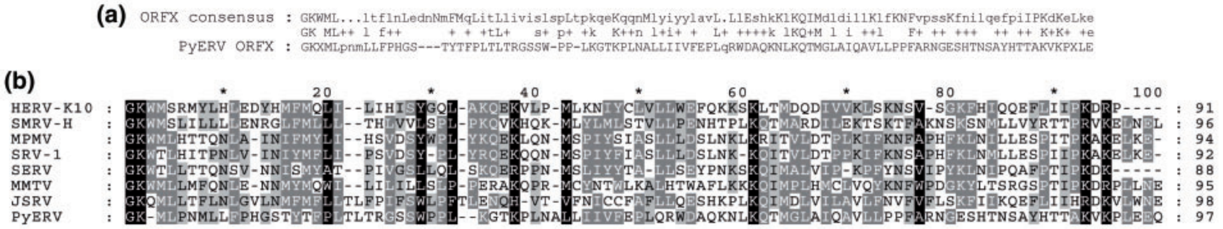
(a)
ORFX consensus : GKWML...ltflnLednNmFMqLitLlivislspLtpkqeKqqnMlyiyylavL.LlEshkKlKQIMdldillKlfKNFvpssKfnilqefpiIPKdKeLke
GK ML++ l f++ + + +tL+ s+ p+ +k K++n l+i+ + L+ ++++k lKQ+M l i ++l F+ ++ +++ ++ ++ K+K+ +e
PyERV ORFX : GKXMLpnmLLFPHGS---TYTFPLTLTRGSSW-PP-LKGTKPLNALLIIVFEPLqRWDAQKNLKQTMGLAIQAVLLPPFARNGESHTNSAYHTTAKVKPXLE
(b)
* 20 * 40 * 60 * 80 * 100
HERV-K10 : GKWMSRMYLHLEDYHMFMQLI--LIHISYGQL-AKQEKVLP-MLKNIYCLVLLWEFQKKSKLTMDQDIVVKLSKNSV-SGKFHIQQEFLIIPKDRP---- : 91
SMRV-H : GKWMSLILLLLENRGLFMLLL--THLVVLSPL-PKQVKHQK-MLYLMLSTVLLPENHTPLKQTMARDILEKTSKTFAKNSKSNMLLVYRTTPRVKELNEL : 96
MPMV : GKWMLHTTQNLA-INIFMYLI--HSVDSYWPL-YKQEKLQN-MSPIYSIASLLLDSLNKLKRITVLDTPLKIFKNFAPHFKLNILLESPITPKAKELKE- : 94
SRV-1 : GKWTLHITPNLV-INIYMFLI--PSVDSY-PL-YRQEKQQN-MSPIYFIASLLLDSLNK-KQITVLDTPPKIFKNSAPHFKLNMLLESPIIPKAKELKE- : 92
SERV : GKWTLLTTQNSV-NNISMYAT--PIVGSLLQL-SKQERPPN-MSLIYYTA-LLSEYPNKSKQIMALVIP-KPFYNSVIPYKLNIPQAFPTIPKDK----- : 88
MMTV : GKWMLLMFQNLE-NNMYMQWI--LILILLSLP-PERAKQPR-MCYNTWLKALHTWAFLKKKQIMPLHMCLVQYKNFWPDGKYLTSRGSPTIPKDRPLLNE : 95
JSRV : GKQMLLTFLNLGVLNMFMFLLTLFPIFSWLPFTLENQH-VT-VFNICCFAFLLQESHKPLKQIMDLVILAVLFNVFVFLSKFIIKQEFLIIHRDKVLWNE : 98
PyERV : GK-MLPNMLLFPHGSTYTFPLTLTRGSSWPPL--KGTKPLNALLIIVFEPLQRWDAQKNLKQTMGLAIQAVLLPPFARNGESHTNSAYHTTAKVKPLEEQ : 97

Figure 4. (**a**) Pairwise alignment between the ORFX MRC sequence and the PyERV–ORF X. (**b**) Multiple alignment.

to a portion of the mammalian adenosine receptor subtype 3 (60). It is still unclear if this ORF is functional (it shows several stop codons in other betaretroviruses), but it is well preserved in both endogenous and exogenous JRSV isolates (61), and we have also found this ORF to be present in other betaretroviruses characteristic of humans, primates and mice, as shown in Figure 4b. We therefore confirm that ORF-X is at least a feature specific of almost all betaretroviruses (another question is if this ORF is functional indeed). With this and based on the significant degree of sequence similarity displayed by PyERV to betaretroviruses, as well as on their identical *gag-dut/pro-pol-env* plus ORF-X organization, we may definitively conclude that PyERV is pure and exclusively a betaretrovirus and likely a D-type betaretrovirus. However, a very interesting point arises from this analysis because if PyERV is a true recombinant, then the simplest hypothesis to explain the emergence of D-type betaretroviruses is that the recombination event between gammaretroviruses and B-type betaretroviruses is more ancient than previously thought. The debate is open.

CONCLUDING REMARKS

The GyDB project pursues the fascinating goal of analyzing and classifying the non-redundant diversity of mobile genetic elements in the context of the Tree of Life, and based on their evolutionary profiles. Due to their impressive molecular diversity, the GyDB is a long-term project that has been arranged in a database in continuous progress, and must be achieved in stages. In this first version, we contemplate the eukaryotic *Ty3/Gypsy* and *Retroviridae* LTR retroelements and demonstrate that the entire molecular diversity inherent to these two groups of LTR retroelements may be used as a main criterion of classification to generate a comprehensive collection of molecular profiles and phylogenies. We pay special attention to non-redundant elements displaying the full-length genome available and a certain degree of distance, as well as to how their entire coding product may be collectively aligned or related in terms of protein domain architecture with other lineages and elements. This is an effort worth making, as we have been able to infer the evolutionary perspectives of the elements we classify based on the complete internal region they commonly display. The GyDB is thus a small but highly informative database established within a phylogenetic context of classification, useful in viral taxonomy and capable of facilitating further identification and analysis of new LTR retroelement species. However, the most captivating aspect of our project is that we dedicate a share of our efforts to the interpretation of our analyses. In Llorens and Moya (manuscript submitted for publication, PLoS ONE) we differentiate the entire clan AA in monophyletic groups of homodomain peptidases in order to reconstruct the ancestral state for each monophyletic group and a consensus template that approximates the molecular phenotype of an ancestor from which the entire clan AA evolves. In another forthcoming study (in preparation) we phylogenetically and comparatively explore the evolutionary meaning of gag-pro-pol diversity. Following from our results, we introduce a guiding principle—the Three Kings Hypothesis—with which we suggest that the early origins of the *Retroviridae* diversity might be more ancient than previously thought, and polyphyletic. We will incorporate in the next GyDB version new non-redundant elements belonging to other LTR retroelement lineages. We think all these incorporations will allow the GyDB to enable exciting insights, leading to a better understanding of the taxonomy and evolutionary history of LTR retroelements. However, as the annotation of new *Ty3/Gypsy* and *Retroviridae* lineages (25,62–64) is constantly growing and we may have not considered in this version, sequences phylogenetically relevant to the database background, the *Ty3/Gypsy* and *Retroviridae* scenario is always open for further evidence. The GyDB project is freely available at http://gydb.uv.es.

ACKNOWLEDGEMENTS

We would like to thank Rachel Epstein for editorial revision, Joaquin Panadero and Miguel Vicente Ripollés for their collaboration in this project, and two anonymous reviewers for useful comments improving this manuscript and the database background. We are especially grateful to the Servei Central de Suport a la Investigació Experimental of the University of Valencia, for technical support, and to all contributors detailed in a list available at http://biotechvana/loader.php?page = policy_gydb. The GyDB project has been awarded the NOVA 2006 by IMPIVA and Conselleria d`Empresa, Universitat I Cìencia of Valencia. The research has been partly supported by European Union funding grants IMCBTA/2005/45, IMIDTD/2006/158 and IMIDTD/2007/33 from IMPIVA, and by grant BFU2005-00503 from MEC to A.M. Funding to pay the Open Access publication charges for this article was provided by University of Valencia.

Conflict of interest statement. None declared.

REFERENCES

1. McClintock,B. (1948) Mutable loci in maize. *Carnegie Inst. Wash. Year book*, **47**, 155–169.
2. Kazazian,H.H.Jr (2004) Mobile elements: drivers of genome evolution. *Science*, **303**, 1626–1632.
3. Temin,H.M. (1989) Reverse transcriptases. Retrons in bacteria. *Nature*, **339**, 254–255.
4. Craig,N.L., Craigie,R., Gellert,M. and Lambowitz,A.M. (2002) *Mobile DNA II*. ASM Press, Washington, DC.
5. Mizuuchi,K. (1992) Transpositional recombination: mechanistic insights from studies of mu and other elements. *Annu. Rev. Biochem.*, **61**, 1011–1051.
6. Wessler,S.R., Bureau,T.E. and White,S.E. (1995) LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. *Curr. Opin. Genet. Dev.*, **5**, 814–821.
7. Bureau,T.E., Ronald,P.C. and Wessler,S.R. (1996) A computer-based systematic survey reveals the predominance of small inverted-repeat elements in wild-type rice genes. *Proc. Natl Acad. Sci. USA*, **93**, 8524–8529.
8. Lynch,M. and Conery,J.S. (2003) The origins of genome complexity. *Science*, **302**, 1401–1404.
9. Poiesz,B.J., Ruscetti,F.W., Gazdar,A.F., Bunn,P.A., Minna,J.D. and Gallo,R.C. (1980) Detection and isolation of type C retrovirus particles from fresh and cultured lymphocytes of a patient with cutaneous T-cell lymphoma. *Proc. Natl Acad. Sci. USA*, **77**, 7415–7419.
10. Yoshida,M., Miyoshi,I. and Hinuma,Y. (1982) Isolation and characterization of retrovirus from cell lines of human adult T-cell leukemia and its implication in the disease. *Proc. Natl Acad. Sci. USA*, **79**, 2031–2035.
11. Barre-Sinoussi,F., Chermann,J.C., Rey,F., Nugeyre,M.T., Chamaret,S., Gruest,J., Dauguet,C., xler-Blin,C., Vezinet-Brun,F. *et al.* (1983) Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science*, **220**, 868–871.
12. Gallo,R.C., Salahuddin,S.Z., Popovic,M., Shearer,G.M., Kaplan,M., Haynes,B.F., Palker,T.J., Redfield,R., Oleske,J. *et al.* (1984) Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS. *Science*, **224**, 500–503.
13. Levy,J.A. and Shimabukuro,J. (1985) Recovery of AIDS-associated retroviruses from patients with AIDS or AIDS-related conditions and from clinically healthy individuals. *J. Infect. Dis.*, **152**, 734–738.
14. Edwards,C.M., Edwards,S.J., Bhumbra,R.P. and Chowdhury,T.A. (2003) Severe refractory hypercalcaemia in HTLV-1 infection. *J. R. Soc. Med.*, **96**, 126–127.
15. UNAIDS. (2007) 06 Annual Report. *Making the Money Work*. UNAIDS, WHO, Geneva.
16. Saigo,K., Kugimiya,W., Matsuo,Y., Inouye,S., Yoshioka,K. and Yuki,S. (1984) Identification of the coding sequence for a reverse transcriptase-like enzyme in a transposable genetic element in *Drosophila melanogaster*. *Nature*, **312**, 659–661.
17. Mount,S.M. and Rubin,G.M. (1985) Complete nucleotide sequence of the Drosophila transposable element copia: homology between copia and retroviral proteins. *Mol. Cell. Biol.*, **5**, 1630–1638.
18. Clare,J. and Farabaugh,P. (1985) Nucleotide sequence of a yeast Ty element: evidence for an unusual mechanism of gene expression. *Proc. Natl Acad. Sci. USA*, **82**, 2829–2833.
19. Eickbush,T.H. (1994) Origin and evolutionary relationships of LTR retroelements. In Morse,S.S. (ed), *The Evolutionary Biology of Viruses*. Raven, New York, pp. 121–157.
20. Kim,A., Terzian,C., Santamaria,P., Pelisson,A., Purd'homme,N. and Bucheton,A. (1994) Retroviruses in invertebrates: the gypsy retrotransposon is apparently an infectious retrovirus of *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA*, **91**, 1285–1289.
21. Song,S.U., Gerasimova,T., Kurkulos,M., Boeke,J.D. and Corces,V.G. (1994) An env-like protein encoded by a Drosophila retroelement: evidence that gypsy is an infectious retrovirus. *Genes Dev.*, **8**, 2046–2057.
22. Pantazidis,A., Labrador,M. and Fontdevila,A. (1999) The retrotransposon Osvaldo from *Drosophila buzzatii* displays all structural features of a functional retrovirus. *Mol. Biol. Evol.*, **16**, 909–921.
23. Bowen,N.J. and McDonald,J.F. (1999) Genomic analysis of *Caenorhabditis elegans* reveals ancient families of retroviral-like elements. *Genome Res.*, **9**, 924–935.
24. Wright,D.A. and Voytas,D.F. (2002) Athila4 of Arabidopsis and Calypso of soybean define a lineage of endogenous plant retroviruses. *Genome Res.*, **12**, 122–131.
25. Volff,J.N., Lehrach,H., Reinhardt,R. and Chourrout,D. (2004) Retroelement dynamics and a novel type of chordate retrovirus-like element in the miniature genome of the tunicate *Oikopleura dioica*. *Mol. Biol. Evol.*, **21**, 2022–2033.
26. Wright,D.A. and Voytas,D.F. (1998) Potential retroviruses in plants: Tat1 is related to a group of Arabidopsis thaliana Ty3/gypsy retrotransposons that encode envelope-like proteins. *Genetics*, **149**, 703–715.
27. Coffin,J.M., Huges,S.H. and Varmus,H.E. (1997) *Retroviruses*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
28. Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G. (1997) The CLUSTAL_X windows interface: flexible

strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.*, **25**, 4876–4882.

29. Eickbush,T.H. and Malik,H.S. (2002) Origin and evolution of retrotransposons. In Craig,N.L., Craigie,R., Gellert,M. and Lambowitz,A.M. (eds), *Mobile DNA II*. ASM Press, Washington, DC, pp. 1111–1144.
30. Pearl,L. and Blundell,T. (1984) The active site of aspartic proteinases. *FEBS Lett.*, **174**, 96–101.
31. Malik,H.S. and Eickbush,T.H. (1999) Modular evolution of the integrase domain in the Ty3/Gypsy class of LTR retrotransposons. *J. Virol.*, **73**, 5186–5190.
32. Xiong,Y. and Eickbush,T.H. (1990) Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J.*, **9**, 3353–3362.
33. Marin,I. and Llorens,C. (2000) Ty3/Gypsy retrotransposons: description of new *Arabidopsis thaliana* elements and evolutionary perspectives derived from comparative genomic data. *Mol. Biol. Evol.*, **17**, 1040–1049.
34. Gorinsek,B., Gubensek,F. and Kordis,D. (2004) Evolutionary genomics of chromoviruses in eukaryotes. *Mol. Biol. Evol.*, **21**, 781–798.
35. Bae,Y.A., Moon,S.Y., Kong,Y., Cho,S.Y. and Rhyu,M.G. (2001) CsRn1, a novel active retrotransposon in a parasitic trematode, *Clonorchis sinensis*, discloses a new phylogenetic clade of Ty3/gypsy-like LTR retrotransposons. *Mol. Biol. Evol.*, **18**, 1474–1483.
36. Boeke,J.D., Eickbush,T.H., Sandmeyer,S. and Voytas,D.F. (1999) Metaviridae. In Murphy,F.A. (ed), *Virus Taxonomy*. ICTV VIIth report. Springer-Verlag, New York.
37. Hull,R. (1999) Classification of reverse transcribing elements: a discussion document. *Arch. Virol.*, **144**, 209–214.
38. Pringle,C.R. (1999) Virus taxonomy, the universal system of virus taxonomy, updated to include the new proposals ratified by the International Committee on Taxonomy of Viruses during 1998. *Arch. Virol.*, **144**, 421–429.
39. Britten,R.J. (1995) Active gypsy/Ty3 retrotransposons or retroviruses in *Caenorhabditis elegans*. *Proc. Natl Acad. Sci. USA*, **92**, 599–601.
40. Van Regenmortel,M.H.V., Fauquet,C.M., Bishop,D.H.L., Carstens,E.B., Estes,M.K., Lemon,S.M., Maniloff,J., Mayo,M.A., McGeoch,D.J. *et al.* (2000) Virus taxonomy: the classification and nomenclature of viruses. *Seventh Report of the International Committee on Taxonomy of Viruses*. Academia Press, San Diego, California.
41. Wilkinson,D.A., Mager,D.L. and Leong,J.A. (1994) Endogenous human retroviruses. In Levy,J.A. (ed), *The Retroviridae*. Plenum Press, New York, pp. 465–535.
42. International Human Genome Consortium. (2002) Initial sequencing and analysis of the human genome. *Nature*, **420**, 520–562.
43. International Human Genome Consortium. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
44. Gifford,R., Kabat,P., Martin,J., Lynch,C. and Tristem,M. (2005) Evolution and distribution of class II-related endogenous retroviruses. *J. Virol.*, **79**, 6478–6486.
45. Gifford,R. and Tristem,M. (2003) The evolution, distribution and diversity of endogenous retroviruses. *Virus Genes*, **26**, 291–315.
46. Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.
47. Eck,R.V. and Dayhoff,M.O. (1966) *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Silver Spring, Maryland.
48. Kluge,A.G. and Farris,J.S. (1969) Quantitative phyletics and the evolution of anurans. *System Zool*, **18**, 1–32.
49. Margus,T. and McMorris,F.R. (1981) Consensus n-trees. *Bull. Math. Biol.*, **43**, 239–244.
50. Koonin,E.V., Zhou,S. and Lucchesi,J.C. (1995) The chromo superfamily: new members, duplication of the chromo domain and possible role in delivering transcription regulators to chromatin. *Nucleic Acids Res.*, **23**, 4229–4233.
51. Llorens,C. and Marin,I. (2001) A mammalian gene evolved from the integrase domain of an LTR retrotransposon. *Mol. Biol. Evol.*, **18**, 1597–1600.
52. Rawlings,N.D. and Barrett,A.J. (1995) Families of aspartic peptidases, and those of unknown catalytic mechanism. *Methods Enzymol.*, **248**, 105–120.
53. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
54. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
55. Huder,J.B., Boni,J., Hatt,J.M., Soldati,G., Lutz,H. and Schupbach,J. (2002) Identification and characterization of two closely related unclassifiable endogenous retroviruses in pythons (*Python molurus* and *Python curtus*). *J. Virol.*, **76**, 7607–7615.
56. Chatterjee,S. and Hunter,E. (1980) Fusion of normal primate cells: a common biological property of the D-type retroviruses. *Virology*, **107**, 100–108.
57. Sommerfelt,M.A. and Weiss,R.A. (1990) Receptor interference groups of 20 retroviruses plating on human cells. *Virology*, **176**, 58–69.
58. Sonigo,P., Barker,C., Hunter,E. and Wain-Hobson,S. (1986) Nucleotide sequence of Mason-Pfizer monkey virus: an immunosuppressive D-type retrovirus. *Cell*, **45**, 375–385.
59. Elder,J.H., Lerner,D.L., Hasselkus-Light,C.S., Fontenot,D.J., Hunter,E., Luciw,P.A., Montelaro,R.C. and Phillips,T.R. (1992) Distinct subsets of retroviruses encode dUTPase. *J. Virol.*, **66**, 1791–1794.
60. Bai,J., Bishop,J.V., Carlson,J.O. and DeMartini,J.C. (1999) Sequence comparison of JSRV with endogenous proviruses: envelope genotypes and a novel ORF with similarity to a G-protein-coupled receptor. *Virology*, **258**, 333–343.
61. Rosati,S., Pittau,M., Alberti,A., Pozzi,S., York,D.F., Sharp,J.M. and Palmarini,M. (2000) An accessory open reading frame (orf-x) of jaagsiekte sheep retrovirus is conserved between different virus isolates. *Virus Res.*, **66**, 109–116.
62. Quesneville,H., Bergman,C.M., Andrieu,O., Autard,D., Nouaud,D., Ashburner,M. and Anxolabehere,D. (2005) Combined evidence annotation of transposable elements in genome sequences. *PLoS. Comput. Biol.*, **1**, 166–175.
63. Goodwin,T.J. and Poulter,R.T. (2002) A group of deuterostome Ty3/gypsy-like retrotransposons with Ty1/copia-like pol-domain orders. *Mol. Genet. Genomics*, **267**, 481–491.
64. Gladyshev,E.A., Meselson,M. and Arkhipova,I.R. (2007) A deep-branching clade of retrovirus-like retrotransposons in bdelloid rotifers. *Gene*, **390**, 136–145.