# The GyDB collection: Ty3/Gypsy and Retroviridae LTR retroelements and related nonviral proteins

Lloréns, C.[1,2] Futami, R.[1] and Moya, A.[2,3]

1- Biotechvana, Valencia, Spain
2- Instituto Cavanilles de Biodiversitat i Biología Evolutiva, Universitat de València, Spain
3- CIBER de Epidemiología y Salud Pública (CIBERESP), Spain

Corresponding author: carlos.llorens@biotechvana.com

In this paper we introduce the GyDB collection, a non-redundant compilation of manually refined multiple alignments, hidden markov model profiles, and majority-rule consensus sequences based on all currently known protein products encoded by Ty3/Gypsy and Retroviridae LTR retroelements and related nonviral proteins. Alignments are available in six formats: Fasta, Pir, Msf, Stockholm, Clustal, Phylip plus a web-shaded HTML format to facilitate preserved motif visualization. The HTML format includes hyperlinks to each sequence's Genbank accession at NCBI. Hidden markov model profiles and majority-rule consensus sequences were constructed based on each protein domain consensus accepted per monophyletic group of LTR retroelements and protein domain.

Keywords: Ty3/Gypsy | Retroviridae | clan AA | GIN-1 | chromodomains

## INTRODUCTION

Bioinformatics and computational biology engage the use of several techniques to understand the gene organization in biological genomes through the characterization of Open Reading Frames (ORFs). Sequencing efforts have revealed that mobile genetic elements are more widely distributed in eukaryotes than previously thought. In an attempt to implement knowledge in this field, we have built the Gypsy Database (GyDB) of mobile genetic elements (1), a research project in which we analyze and classify non-redundant mobile genetic elements based on their evolutionary profiles. The GyDB project is accessible at (URL 2). The first version contemplates the *Ty3/Gypsy* and *Retroviridae* LTR retroelements (retrotransposons and retroviruses) of eukaryotic organisms and several nonviral protein groups related to them.

## OVERVIEW

*Ty3/Gypsy* and *Retroviridae* LTR retroelements are two groups of evolutionarily related mobile genetic elements that reverse-transcribe their RNA genome into a double-stranded DNA copy inserted in the host cell genome. According to the International Committee on the Taxonomy of Viruses (ICTV) (2) and including ERV-L elements, the *Retroviridae* comprise a certain number of lineages. Based on genome complexity the *Retroviridae* may also be divided in two categories, simple and complex. The main difference consists in that simple retroviruses present the same basal LTR-gag-pol-env-LTR genome structure than *Ty3/Gypsy* retroviruses, while complex retroviruses incorporate in their genomes some additional accessory genes, usually needed to adjust diverse aspects of their replication and infectivity. On the other hand, a number of *Ty3/Gypsy* clades and genera have been described in the genomes of plants, fungi, and bilateria organisms (3-15). We provide a detail of these lineages in (1) and the GyDB (URL 2) extends discussions and citations of all retroelement species and lineages we classify.

The basal genome structure of a *Ty3/Gypsy* or *Retroviridae* LTR retroelement consist in an internal region flanked by two normally homologous non-coding DNA sequences known as Long Terminal Repeats (LTRs). A nucleotide sequence of 18 nt in size, used as a Primer Binding Site (PBS) in the retrotranscription process, is found downstream to the 5´LTR. *Ty3/Gypsy* and *Retroviridae* elements usually harbor a Polypurine Tract (PPT) of ~10 A/G, found preceding the 3´LTR. The PPT sequence is responsible for the beginning of the proviral DNA strand (+) synthesis. The internal region contains the ORFs characteristic of LTR retroelements and are usually arranged as follows: a *gag* gene coding for a gag polyprotein precursor containing the matrix (MA), capsid (CA), and nucleocapsid (NC) domains; and a *pol* gene coding for a pol polyprotein precursor usually containing the protease (PR), reverse transcriptase (RT), ribonuclease H (RNAse H), and integrase (INT) domains. PR may be encoded by a gene alone, as a part of the gag polyprotein, or in frame with a dUTPase domain in certain vertebrate retroviruses. As the absence or presence of *env* is the main difference between a LTR retrotransposon and a retrovirus, those LTR retroelements containing a third ORF *env* are considered to be true or potential retroviruses. This ORF normally codifies for an envelope (env) precursor containing the surface (SU) and transmembrane (TM) domains.

PRs encoded by *Ty3/Gypsy,* and *Retroviridae* and other groups of LTR retroelements belong to clan AA of aspartic peptidases (16). Clan AA is a group of proteolytic enzymes that use an aspartate dyad and a molecule of water to hydrolyze a peptide bond (17). Aspartic peptidases belonging to clan AA (CAPs) are structurally divided in two large groups (18;19). The first comprises the nonviral eukaryotic pepsin monomers structured in two protein domains of similar architecture (set of physiochemical properties of residues preserved in a protein domain), the second embraces the homodomain proteases that dimerize in their active form and are usually part of the pol polyprotein encoded by eukaryotic LTR retroelements.

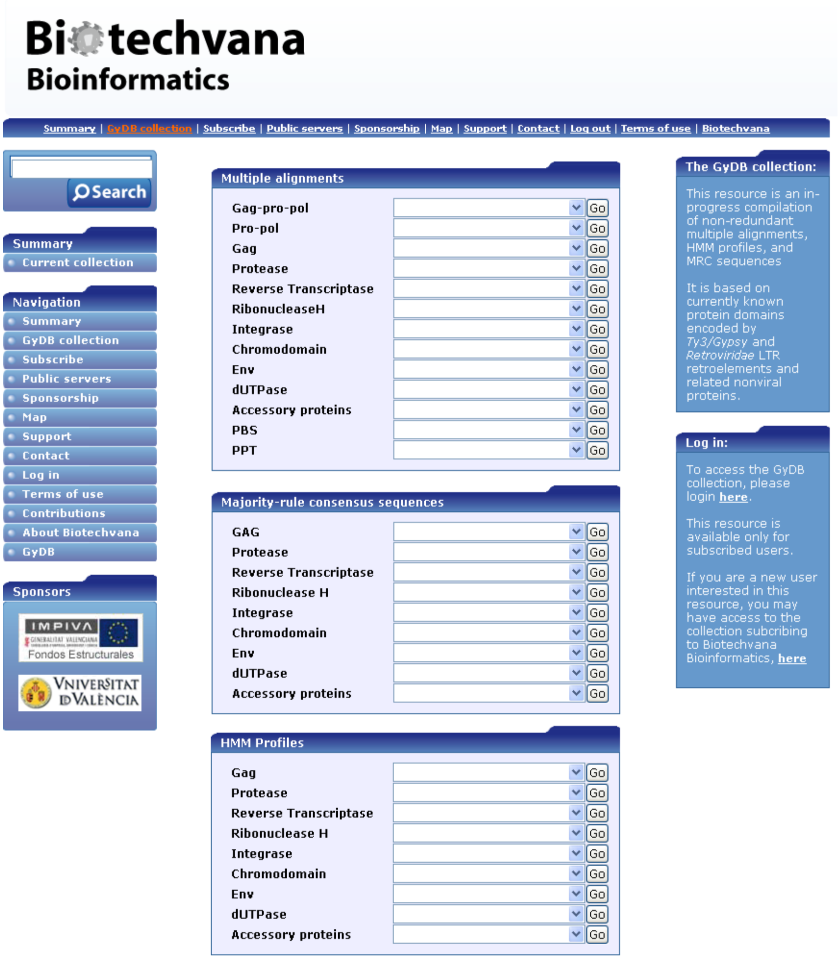
Chromoviruses (7;13-15) constitute an ancient branch of *Ty3/Gypsy* LTR retrotransposons described in genomes of plant, fungi, and vertebrate organisms. Chromoviruses may generally be differentiated from other *Ty3/Gypsy* retrotransposons due to their presentation of a chromodomain at the C-terminal end of their INTs (10). The chromodomain (chromatin organization modifier) is a protein domain of 40–50 amino acids, identified in a variety of proteins (20;21).

GIN-1 INTs are nonviral eukaryotic INTs evolutionarily related to those encoded by *412/mdg1 Ty3/Gypsy* LTR retrotransposons of protostomes (22). This evidence suggests that this gene was likely recruited by certain deuterostome genomes from a *Ty3/Gypsy* retrotransposon along to the evolution.

The GyDB collection contemplates all protein domains encoded by 120 *Ty3/Gypsy* and *Retroviridae* full-length genomes, 307 non-redundant clan AA aspartic peptidases, 111 eukaryotic chromodomains, and 6 GIN-1 integrases; all collected from the National Center of Biotechnology Information (NCBI) (URL 3) and MEROPS (URL 4).  Based on all this material we have built an in-progress collection of multiple alignments and molecular profiles useful in retroelement taxonomy and identification of new retroelement species. The collection is presented via a web server divided in three categories: multiple alignments, Hidden Markov Model  (HMM) profiles, and majority-rule consensus (MRC) sequences. As shown in Figure 1, each category has a set of drop-down lists that display items for multiple alignments, HMM profiles, and MRC sequences grouped by domains.

## Multiple Alignments

*Ty3/Gypsy* and *Retroviridae* multiple alignments were obtained and manually refined using CLUSTAL X (23) and GENEDOC (URL 5) based on the DNA sequences and protein domains summarized in Table 1. This table contemplates two multiple alignments based on the *Ty3/Gypsy/Retroviridae* PBS and PPT motifs, six multiple alignments based on single protein domain motifs, and four multiple alignments constructed from the concatenation of two or more protein domains. Alignments are available in six formats: Fasta, Pir, Msf, Stockholm, Clustal, Phylip plus a web-shaded HTML format to facilitate preserved motif visualization. The HTML format includes hyperlinks to each sequence's Genbank accession at NCBI. The following groups of amino acid similarity were used in the protein alignment refinement; [T,S small nucleophile amino acids] [K,R,H basic amino acids], [D,E,N,Q acidic amino acid and relative amides], and [L,I,V,M,A,G,P,F,Y,W hydrophobic amino acids]. These groups of similarity take into account the physiochemical properties of amino acids, providing, by way of the sequence editor's shaded mode, an amino acid architecturally improved visualization for conserved protein domains, and much better visualization for non-conserved protein do-



This resource is an in-progress compilation of non-redundant multiple alignments, HMM profiles, and MRC sequences

It is based on currently known protein domains encoded by *Ty3/Gypsy* and *Retroviridae* LTR retroelements and related nonviral proteins.

To access the GyDB collection, please login **here**.

This resource is available only for subscribed users.

If you are a new user interested in this resource, you may have access to the collection subcribing to Biotechvana Bioinformatics, **here**

Figure 1: Screenshot of the GyDB Collection web site.

Table 1: *Ty3/Gypsy* and *Retroviridae* multiple alignments

| Alignment | Type | Composed by |
|---|---|---|
| Gag-pro-pol | Protein | CA-NC-PR-RT-RNAseH-INT |
| Pro-pol | Protein | PR-RT-RNAseH-INT |
| Gag | Protein | CA-NC |
| Protease | Protein | PR |
| RT | Protein | RT |
| RNAse H | Protein | RNAse H |
| INT | Protein | INT |
| Chromodomain | Protein | CHR |
| Env | Protein | SU-TM |
| dUTPase | Protein | dUTPase |
| Primer Binding Site (PBS) | DNA | PBS |
| Polypurine Tract (PPT) | DNA | PPT |

mains. Protein domains were automatically aligned and refined taking into consideration the cores described for RT in (3;7;10), and the RNAse H and INT in (10;22,24,25). Due to their difficulty, the gag polyprotein, the PR domain and clan AA, and the env polyprotein, were aligned as follows:

Due to the low degree of preservation, analyses performed based on the gag polyprotein are rarely reported. We dismissed MA from the analysis by its extreme variability but the CA-NC region revealed two zones of similarity among Ty3/Gypsy and Retroviridae gags; the major homology region (MHR) of CA (26) and the zinc finger Cys-X2-Cys-X4-His-X4-Cys (CCHC) array found at NC (27). We compared gag sequences against themselves through a number of BLAST (28) searches against the CORES database available at GyDB. The best hits of similarity were consistent with the clades and genera reported in prior studies (2;3;7-13;29). This means that in spite of the fast rate of evolution, gag sequences encoded by all LTR retroelements belonging to a clade or to a genus are more similar among them than to other sequences (data not shown). We used this information to obtain a number of CA-NC alignments, each based on a clade or a genus. BLAST analyses also revealed gag similarity among sequences belonging to different clades and genera (data not shown). We used the range of similarities among clades to manually join all alignment in just a single alignment, which was refined through a number of manual lineage-to-lineage comparisons using the sequence editor.

Ty3/Gypsy and Retroviridae PRs present identical or higher difficulty than that of the gag polyprotein. It is known, however, that together with several monophyletic groups of nonviral proteolytic enzymes, all currently known LTR retroelement PRs belong to clan AA of aspartic peptidases (CAPs). At the primary structure level, the most prominent phenotype of almost but not all CAPs is the catalytic DT/SG triad (30) displayed near to the N-terminus, and a glycine close to the C-terminus that is preceded by two hydrophobic residues (normally isoleucine and leucine) (31). The presence of these two DTG and ILG motifs does not necessarily denote CAP function, but the most conserved part (core) of retropepsins (in this work, all CAPs encoded by vertebrate retroviruses except spumaretroviruses) is usually a useful control to identify new CAPs. According to MEROPS (16) clan AA includes a number of families with a clan sub-letter of family classification assigned. As summarized in Table 2, we differentiated the clan in 38 monophyletic clusters according to previously known antecedents based on MEROPS insights and LTR retroelement evolution. For instance, the pepsin monomer consists in two similar CAP domains we have divided in two homodomains (pepsin domain 1 and pepsin domain 2) to be independently aligned. However, as pepsins split at MEROPS in two subfamilies (A1A and A1B), we finally consider four clusters - A1aD1 and A1aD2 and A1bD1 and A1bD2 - for homodomains 1 and 2 of subfamilies A1A and A1B, respectively. Table 2 also assumes a number of clusters based on *Ty3/Gypsy* and *Retroviridae* phylogenies (1) and following MEROPS, two additional clusters based on the CAPs encoded by *Caulimoviridae* and *Pseudoviridae* (*Ty1/Copia*) LTR retroelements. Finally, Table 2 considers other clusters not classified at MEROPS based on the CAPs encoded by *Bel* retroelements (11), the CAP of the *Bs-1* LTR retrotransposon (32), and a number of homodomain nonviral CAPs (HNCAPs) described in prokaryotes (COG3577 and COG5550) and eukaryotes (DDI, NIX) (33,34). Little is known about prokaryotic HNCAPs, they are single ORFs encoding for a protein domain of nearly 120 residues in size widely distributed in proteobacteria and having at least one known representative in archaea (see the clan AA tree in the Section "Phylogenies" at GyDB (URL 6). DDI proteins have been more extensively studied. DDI enzymes are clearly related with NIX sequences and are widely distributed in the genomes of plants, fungi and animals (33). In contrast, the availability of characterized NIX sequences suggests at present that *nix* genes are only found in vertebrate genomes. DDI proteins exhibit a central CAP domain usually accompanied by one of two ubiquitin domains not displayed in NIX proteins (33;34). The recent characterization of the *S. cerevisiae* DDI-like X-ray crystal structure corroborates that DDI is a dimer with similar fold than retroviral CAPs (35). Eukaryotic HNCAPs gained interest with the recently reported specific expression of another set of eukaryotic nonviral proteins (SASPases) in human and mouse epidermis (36,37). SASPases are carriers of a CAP homodomain similar to that displayed in DDI enzymes, and have been related with side effects on skin induced by protease inhibitors in anti-AIDS therapy (37). To align the set of retrieved clan sequences, we followed the clustering differentiation summarized in Table 2 established one alignment based on each cluster with the exception of those based on a single sequence. Of the set, we selected the lentiviral alignment to use the well known retropepsin core as a template to align the remaining sequences, alignment-by-alignment. The retropepsin core is based on a poorly preserved hydrophobic architecture of ~90-150 residues in size (31,38,39). We then introduced gaps where needed and refined the resultant alignment through multiple comparisons among all possible combinations of clusters. We refined this alignment through the identification of an ancestral consensus common to all CAPs we have called DTG/ILG template. We provide extensive details regarding this template and clan AA relationships in a forthcoming manuscript in preparation.

The GyDB collection also contemplates two compound alignments: gag-pro-pol, and pro-pol (for simplicity, "pol") constructed from the concatenation of all protein products encoded by the gag-pro-pol region of *Ty3/Gypsy* and *Retroviridae* LTR retroelements. Here, we consider the PR domain to be another pol component, as it is usually part of the polyprotein and has a phylogenetic signal that is low, yet similar to that of other pol protein domains (see PR tree in the "Section Phylogenies" at GyDB (URL 7)). The pol alignment concatenates the PR, RT, RNAse H, and INT pol polyprotein domains in a single multiple alignment, from the catalytic "DTG" PR triad (30) to the GPY/F module (10). To do this, we used a

Table 2. Clan AA clusters

| Taxonomy | Clusters | Seq | Hosts | MEROPS |
|---|---|---|---|---|
| Retroviridae | Lentiviridae | 11 | Amniota | A2A |
| | Alpharetroviridae | 3 | Amniota | A2A |
| | Betaretroviridae | 8 | Amniota | A2A |
| | Gammaretroviridae | 13 | Amniota | A2A |
| | Deltaretroviridae | 4 | Amniota | A2A |
| | Epsilonretroviridae | 1 | Amniota | A2A |
| | Spumaretroviridae | 6 | Amniota | A9 |
| | MuERV-L | 1 | Amniota | A2A |
| Ty3/Gypsy | 412/mdg1 | 2 | Arthropoda | No family |
| | Athila | 9 | Viridiplantae | No family |
| | Cer1 | 1 | Nematoda | No family |
| | Cer2-3 | 2 | Nematoda | No family |
| | Chrofung* | 14 | Fungi/vertebrates | No family |
| | CsRN1 | 2 | Protostomia | No family |
| | CRM | 3 | Viridiplantae | No family |
| | Del | 7 | Viridiplantae | No family |
| | Errantiviridae | 14 | Arthropoda | A2C and 2G |
| | Galadriel | 3 | Viridiplantae | No family |
| | Mag | 7 | Bilateria | No family |
| | Micropia/mdg3 | 3 | Arthropoda | No family |
| | Osvaldo | 4 | Arthropoda | A2D |
| | Reina | 4 | Viridiplantae | No family |
| | Tat | 10 | Viridiplantae | No family |
| | TF1-2 | 2 | Fungi | A2E |
| | Ty3 | 3 | Fungi | A2B |
| Other retroelement | Bel | 16 | Bilateria | No family |
| | Bs-1 | 1 | Viridiplantae | No family |
| | Caulimoviridae | 18 | Viridiplantae | A3 |
| | Ty1/Copia | 26 | Eukaryota | A11 |
| Non-viral prokaryotic | COG5550 | 10 | Prokaryota | No family |
| | COG3577 | 20 | Prokaryota | No family |
| Non-viral eukaryotic | DDI | 20 | Eukaryota | No family |
| | NIX-1 | 5 | Amniota | No family |
| | SASPases | 6 | Amniota | No family |
| Pepsin domains | Pepsin_A1a | 27 | Eukaryota | A1A |
| | Pepsin_A1b | 5 | Eukaryota | A1B |

*We have used the descriptor "Chrofung" to describe a cluster based fungi and vertebrate chromoviruses (for more information regarding chromoviruses see URL 8)

PHP script named Joint Alignments Server (See the Section "Scripts" in Biotechvana Bioinformatics). The pol alignment was again concatenated with the CA-NC alignment obtained based on the gag polyprotein using JAS to have a single multiple alignment of nearly 1800 residues in size. This alignment describes the entire protein product encoded by the gag-pol internal region, from the CA to the GPY/F module. The phylogenetic tree inferred based on the gag-pol alignment is available in the Section phylogenies at GyDB (URL 8) and it is the main criterion of clustering classification we currently follow at GyDB (see (1)).

Little is known about what is common at the primary structure level among retroviral env polyproteins. However, a conserved amino acidic "KRG" motif (also termed "R-X-K-R") has been described preceding a zone common to all retroviral env-like polyproteins (40-43). This motif is the consensus cleavage site recognized by the cellular endopeptidase that cleaves the env precursor into the SU and TM peptides. Also, the template "R-x(2)-R-X(5,6)-[GE]-x(5)-[LV]-x-Gx(2)-D-x(2)-D" has been suggested for the "*in sílico*" detection of insect retroviral env sequences in databanks (36). This information was used to obtain env multiple alignments based on *Athila* elements (9), errantiviruses (44), and vertebrate retroviruses (2).

"Accessory protein" multiple alignments were performed based on the detail of *Retroviridae* accessory genes provided in Llorens et al (1) (see also Table 3 in this paper). A discussion of all accessory genes is also available at GyDB (URL 9).

## Hidden Markov Model Profiles

The identification of consensus sequences facilitates the identification of relationships and taxonomy of sequences, as well as the discernment of conserved motifs that may be characteristic of protein domains. The detection of well-defined protein motifs or domains is a useful tool to classify proteins into families and these classifications can be used to assign putative physiological roles to new discovered proteins (45). One of the most powerful methodologies in this area is provided by Profile Hidden Markov Models (HMM) (46), which are statistical models constructed from multiple sequence alignments. HMM profiles capture position-specific information on the degree of conservation of residues in each column of the alignment. Taking into account the monophyletic clusters reported by gag-pol phylogenetic analyses performed at GyDB (URL 8), we have constructed a comprehensive collection of 155 HMM profiles using HMMER 2.3.2 (URL 10). As shown in Table 3, we divided the alignments we summarize in Table 1 based on single protein domains (RT, RNAse H, etc) in monophyletic group sub-alignments to construct an HMM profile based on each *Ty3/Gypsy* and *Retroviridae* clade and genera (for more details regarding *Ty3/Gypsy* and *Retroviridae* monophyletic groups see (1)). We also constructed additional HMM profiles from GIN-1 and chromodomains alignments, and finally reconstructed and profiled the ancestral consensus of each clan AA cluster summarized in Table 2 (we give more details regarding this strategy in a forthcoming manuscript in preparation).

### Majority-rule consensus sequences

The majority-rule consensus (MRC) sequence methodology consists in the creation of a single consensus from a set of sequences. In MRC sequences, highly conserved residues (probability greater than or equal to 90 percent for DNA and greater than or equal to 50 percent for protein) are shown in uppercase, and less conserved residues in lowercase. We used the collection of HMM profiles to construct a derived collection of 155 MRC sequences using HMMER (URL 10) .

## CONCLUDING REMARKS

The GyDB collection contemplates by monophyletic groups, all the protein products encoded by *Ty3/Gypsy* and *Retroviridae* LTR retroelements and several groups of related nonviral proteins. Due to high divergence, many of the alignments we provide have been manually constructed and therefore are a highly informative set of tools. The collection is in continuous progress and runs parallel to the GyDB Project. Note that the evidence of new *Ty3/Gypsy* and *Retroviridae* species and lineages also grows parallel to sequencing projects. Although the current version of the GyDB contemplates a representative subset of the diversity of these two groups, the database requires continuous updates of sequences phylogenetically relevant for the database background. This is however, the most interesting aspect of this collection because it is a significant tool to compare and evaluate each new characterized finding in the area. Reversely, this feedback is useful to us in order to re-build the collection and obtain more accurate profiles of these two (and other) groups of LTR retroelements.

## ACKNOWLEDGMENTS

We thank Rachel Epstein for language revision and the Servei Central de Suport a la Investigació Experimental (SCSIE) at UVEG for technical support. Biotechvana Bioinformatics has been awarded the NOVA 2006 by IMPIVA and Conselleria d`Empresa, Universitat I Cìencia of Valencia. The research has been partly supported by grants IMCBTA/2005/45, IMIDTD/2006/158 and IMIDTD/2007/33 from IMPIVA, and by grant BFU2005-00503 from MEC to AM.

Table 3. HMM profiles and MRC sequences

| **GAG, RT, RMAse H and INT domains** |
|---|
| ***Retroviridae*** |
| · alpharetroviridae |
| · betaretroviridae |
| · gammaretroviridae |
| · deltaretroviridae |
| · lentiviridae |
| · spumaretroviridae |
| ***Ty3/Gypsy*** |
| · 412/mdg1 |
| · athila |
| · cer2-3 |
| · chrofung |
| · ty3 |
| · crm |
| · galadriel |
| · del |
| · reina |
| ·· TF |
| · errantiviridae |
| · mag |
| · micropia/mdg3 |
| · osvaldo |
| · tat |
| · csrn1 |

Note that we contemplate an HMM profile and an MRC sequence per each protein domain

Table 3. HMM profiles and MRC sequences (continuation)

| **ENV polyprotein** |
|---|
| ***Retroviridae*** |
| · ENV_retroviridae (local *Retroviridae* consensus) |
| · ENV_alpharetroviridae |
| · ENV_betaretroviridae |
| ·ENV_gammaretroviridae |
| · ENV_deltaretroviridae |
| · ENV_lentiviridae |
| ***Ty3/Gypsy*** |
| · ENV_athila |
| · ENV_ errantiviridae |

| **dUTPase domain** |
|---|
| · DUT_retroviridae (Retroviridae dUTPase) |
| · DUT_betaretroviridae |
| · DUT_lentiviridae |

| **Chromodomains** |
|---|
| · CR_all (chromodomains) |
| · CR_shadow(chromoshadow domains) |

| **Non-viral integrases** |
|---|
| · GIN1 |

| ***Retroviridae* accessory proteins** |
|---|
| · BEL1_retroviridae (spumaretroviruses) |
| · BEL2_retroviridae (spumaretroviruses) |
| · BEL3_retroviridae (spumaretroviruses) |
| · NEF_retroviridae (primate lentiviruses) |
| · REV_retroviridae (lentiviruses) |
| · REX_retroviridae (deltaretroviruses) |
| · ROF_retroviridae (deltaretroviruses) |
| · SORF_retroviridae (betaretroviruses SRV-1 and MPMV) |
| · TAT_retroviridae (lentiviruses) |
| · TAX_retroviridae (deltaretroviruses) |
| · TOF_retroviridae (deltaretroviruses) |
| · ORFQ_retroviridae (visna virus-like lentiviruses) |
| · ORFW_retroviridae (visna virus-like lentiviruses) |
| · ORFX_retroviridae (betaretroviruses) |
| · VIF_retroviridae (primate-like lentiviruses) |
| · VIF_Q_retroviridae (lentiviruses) |
| · VPR_retroviridae (primate lentiviruses) |
| · VPX_retroviridae (HIV-2-like lentiviruses) |
| · VPR_VPX_retroviridae (primate lentiviruses) |

Table 3. HMM profiles and MRC sequences (continuation)

| **Clan AA CAPs** |
|---|
| · DTG/ILG_template |
| ***Retroviridae*** |
| · alpharetroviridae |
| · betaretroviridae |
| · gammaretroviridae |
| · deltaretroviridae |
| · lentiviridae |
| · spumaretroviridae |
| ***Ty3/Gypsy*** |
| · 412/mdg1 |
| · athila |
| · cer2-3 |
| · chrofung |
| · ty3 |
| · crm |
| · galadriel |
| · del |
| · reina |
| · TF |
| · errantiviridae |
| · mag |
| · micropia/mdg3 |
| · osvaldo |
| · tat |
| · csrn1 |
| **Other retroelement PRs** |
| · caulimoviruses |
| · ty1/Copia |
| · bel |
| **Nonviral prokaryotic PRs** |
| · cog3577 |
| · cog5550 |
| **Nonviral eukaryotic PRs** |
| · DDI |
| · nix |
| . SASPases |
| · pepsins_A1a |
| · pepsins_A1b |

## LITERATURE

1. Llorens,C., Futami,R., Bezemer,D. and Moya,A. (2008) Nucleic Acids Research (NAR) 36 (Database-Issue):38-46

2. 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., Pringle,C.R. et al. (2000) San Diego, California.

3. Xiong,Y. and Eickbush,T.H. (1990) EMBO J., 9, 3353-3362.

4. Boeke,J.D., Eickbush,T.H., Sandmeyer,S. and Voytas,D.F. (1999) Springer-Verlag, New York.

5. Pringle,C.R. (1999) Archives of Virology, 144, 421-429.

6. Hull,R. (1999) Archives of Virology, 144, 209-214.

7. Marin,I. and Llorens,C. (2000) Mol. Biol. Evol., 17, 1040-1049.

8. Wright,D.A. and Voytas,D.F. (1998) Genetics, 149, 703-715.

9. Wright,D.A. and Voytas,D.F. (2002) Genome Res., 12, 122-131.

10. Malik,H.S. and Eickbush,T.H. (1999) J. Virol., 73, 5186-5190.

11. Bowen,N.J. and McDonald,J.F. (1999) Genome Research, 9, 924-935.

12. Bae,Y.A., Moon,S.Y., Kong,Y., Cho,S.Y. and Rhyu,M.G. (2001) Mol. Biol. Evol., 18, 1474-1483.

13. Gorinsek,B., Gubensek,F. and Kordis,D. (2004) Mol. Biol. Evo., 21, 781-78

14. Gorinsek,B., Gubensek,F. and Kordis,D. (2005) Cytogenet. Genome Res., 110, 543-552.

15. Kordis,D. (2005) Gene, 347, 161-173.

16. Rawlings,N.D., Tolle,D.P. and Barrett,A.J. (2004) Nucleic Acids Research, 32, D160-D164.

17. Fruton,J.S. (1976) Advan. Enzymol., 44, 1-36.

18. Davies,D.R. (1990) Annual Review of Biophysics and Biophysical Chemistry, 19, 189-215.

19. Rawlings,N.D. and Barrett,A.J. (1995) Methods Enzymol., 248, 105-120

20. Koonin,E.V., Zhou,S. and Lucchesi,J.C. (1995) Nucleic Acids Res., 23, 4229-4233.

21. Cavalli,G. and Paro,R. (1998) Curr. Opin. Cell Biol., 10, 354-360.

22. Llorens,C. and Marin,I. (2001) Mol. Biol. Evol., 18, 1597-1600.

23. Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D. G. (1997) Nucleic Acids Res., 25, 4876-4882.

24. Malik,H.S. and Eickbush,T.H. (2001) Genome Res., 11, 1187-1197.

25. Polard, P. and Chandler, M (1995) Mol. Microbiol. 15: 13-23

26. Nakayashiki,H., Matsuo,H., Chuma,I., Ikeda,K., Betsuyaku,S., Kusaba,M., Tosa,Y. and Mayama,S. (2001) Nucleic Acids Res., 29, 4106-4113.

27. Green,L.M. and Berg,J.M. (1989) Proc. Natl. Acad. Sci. U. S. A, 86, 4047-4051.

28. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Nucleic Acids Res., 25, 3389-3402.

29. Britten,R.J. (1995) Proc. Natl. Acad. Sci. U. S. A, 92, 599-601.

30. Pearl,L. and Blundell,T. (1984) FEBS Lett., 174, 96-101

31. Pearl,L.H. and Taylor,W.R. (1987) Nature, 329, 351-354.

32. Johns,M.A., Mottinger,J. and Freeling,M. (1985) EMBO J., 4, 1093-1101.

33. Krylov,D.M. and Koonin,E.V. (2001) Curr. Biol., 11, R584-R587.

34. Puente,X.S., Sanchez,L.M., Overall,C.M. and Lopez-Otin,C. (2003) Nat. Rev. Genet., 4, 544-558.

35. Sirkis,R., Gerst,J.E. and Fass,D. (2006) J. Mol. Biol., 364, 376-387.

36. Matsui,T., Kinoshita-Ida,Y., Hayashi-Kisumi,F., Hata,M., Matsubara,K., Chiba,M., Katahira-Tayama,S., Morita,K., Miyachi,Y. and Tsukita,S. (2006) J. Biol. Chem., 281, 27512-27525.

37. Bernard,D., Mehul,B., Thomas-Collignon,A., Delattre,C., Donovan,M. and Schmidt,R. (2005) J. Invest Dermatol., 125, 278-287.

38. Wlodawer,A. and Gustchina,A. (2000) Biochim. Biophys. Acta, 1477, 16-34.

39. Weber,I.T. (1989) Gene, 85, 565-566.

40. Kim,A., Terzian,C., Santamaria,P., Pelisson,A., Purd'homme,N. and Bucheton,A. (1994 Proc. Natl. Acad. Sci. U. S. A, 91, 1285-1289.

41. Leblanc,P., Desset,S., Dastugue,B. and Vaury,C. (1997) EMBO J., 16, 7521-7531.

42. Lerat,E. and Capy,P. (1999) Mol. Biol. Evol., 16, 1198-1207.

43. Malik,H.S., Henikoff,S. and Eickbush,T.H. (2000) Genome Res., 10, 1307-1318.

44. Eickbush,T.H. and Malik,H.S. (2002) In Craig,N.L., Craigie,R., Gellert,M. and Lambowitz,A.M. (eds.), Mobile DNA II. ASM Press, Washington DC., pp. 1111-1144.

45. Schug,J., Diskin,S., Mazzarelli,J., Brunk,B.P. and Stoeckert,C.J., Jr. (2002) Genome Res., 12, 648-655.

46. Eddy,S.R. (1998) Bioinformatics, 14, 755-763.

## URLS

1. Creative Commons Attribution License: http://creativecommons.org/licenses/by/2.0
2. GyDB: http://gydb.uv.es
3. NCBI: http://www.ncbi.nlm.nih.gov.
4. MEROPS: http://merops.sanger.ac.uk
5. GENEDOC: http://www.psc.edu/biomed/genedoc
6. Clan AA tree: http://gydb.uv.es/gydb/phylogeny.php?tree=clana
7. PR tree: http://gydb.uv.es/gydb/phylogeny.php?tree=pr
8. Gag-pro-pol tree: http://gydb.uv.es/gydb/phylogeny.php?tree=gagpol
9. Accessory proteins: http://gydb.uv.es/gydb/description.php?desc=retroviridae_acc
10. HMMER: http://hmmer.janelia.org

Biotechvana Bioinformatics

## SPONSORS

VNIVERSITAT DE VALÈNCIA

IMPIVA

GENERALITAT VALENCIANA
CONSELLERIA D'EMPRESA, UNIVERSITAT I CIÈNCIA

CEEI VALENCIA

BIOVAL
ASOCIACIÓN DE EMPRESAS BIOTECNOLÓGICAS DE LA COMUNIDAD VALENCIANA

nova06
Premis de l'empresa valenciana

MINISTERIO DE EDUCACIÓN Y CIENCIA

UNIÓN EUROPEA
Fondo Europeo de Desarrollo Regional
Una manera de hacer Europa