D70–D74 *Nucleic Acids Research, 2011, Vol. 39, Database issue* *Published online 29 October 2010*
doi:10.1093/nar/gkq1061

The *Gypsy* Database (GyDB) of mobile genetic elements: release 2.0

Carlos Llorens[1,*], Ricardo Futami[1], Laura Covelli[1], Laura Domínguez-Escribá[1], Jose M. Viu[1], Daniel Tamarit[2], Jose Aguilar-Rodríguez[2], Miguel Vicente-Ripolles[3], Gonzalo Fuster[1], Guillermo P. Bernet[1,4], Florian Maumus[5], Alfonso Munoz-Pomer[1,3], Jose M. Sempere[3], Amparo Latorre[2,6] and Andres Moya[2,6]

[1]Biotechvana, Parc Científic, Universitat de València, Calle Catedrático José Beltrán 2, 46980 Paterna (València), [2]Unidad Mixta de Investigación en Genómica y Salud del Centro Superior de Investigación en Salud Pública (CSISP)-Universitat de València (Instituto Cavanilles de Biodiversidad y Biología Evolutiva), Avenida de Cataluña 21, 46020 València, [3]Departamento de Sistemas Informáticos y Computación (DSIC), Universitat Politècnica de València, Camino de Vera S/N, 46022 València, [4]Instituto Valenciano de Investigaciones Agrarias (IVIA), Carretera Moncada-Naquera, Km 4.5, 46113, Moncada (València), Spain, [5]Institut Jean-Pierre Bourgin, INRA Centre de Versailles-Grignon, Route de Saint-Cyr, 78026 Versailles, France and [6]CIBER en Epidemiologia y Salud Pública (CIBEResp), Parc de Recerca Biomèdica de Barcelona, Calle Doctor Aiguader 88 1[a] Planta, 8003 Barcelona, Spain

Received September 2, 2010; Revised October 11, 2010; Accepted October 13, 2010

ABSTRACT

This article introduces the second release of the Gypsy Database of Mobile Genetic Elements (GyDB 2.0): a research project devoted to the evolutionary dynamics of viruses and transposable elements based on their phylogenetic classification (per lineage and protein domain). The Gypsy Database (GyDB) is a long-term project that is continuously progressing, and that owing to the high molecular diversity of mobile elements requires to be completed in several stages. GyDB 2.0 has been powered with a wiki to allow other researchers participate in the project. The current database stage and scope are long terminal repeats (LTR) retroelements and relatives. GyDB 2.0 is an update based on the analysis of *Ty3/Gypsy*, *Retroviridae*, *Ty1/Copia* and *Bel/Pao* LTR retroelements and the *Caulimoviridae* pararetroviruses of plants. Among other features, in terms of the aforementioned topics, this update adds: (i) a variety of descriptions and reviews distributed in multiple web pages; (ii) protein-based phylogenies, where phylogenetic levels are assigned to distinct classified elements; (iii) a collection of multiple alignments, lineage-specific hidden Markov models and consensus sequences, called GyDB collection; (iv) updated RefSeq databases and BLAST and HMM servers to facilitate sequence characterization of new LTR retroelement and caulimovirus queries; and (v) a bibliographic server. GyDB 2.0 is available at http://gydb.org.

INTRODUCTION

Mobile genetic elements (MGEs) are ubiquitous, autonomous genetic units that often constitute a significant part of their host genomes. It is commonly accepted that mobile DNA elements are powerful vectors for disease and evolution, from which distinct host genes have evolved during the history of life (1,2). The emergence and subsequent role played by viruses and MGEs in the history of life is an exciting topic that requires further investigation. In this respect, researchers aim to discern relevant aspects of the molecular changes responsible for various characteristics in organisms related to horizontal transfer, infection and disease. Among the distinct initiatives launched with the aim of investigating the diversity of MGEs (see for example 3–5) was the Gypsy Database (GyDB) of MGEs (6), a research project devoted to the evolutionary dynamics of viruses and MGEs (and their related host proteins), which was launched in 2008. The GyDB project is a highly

*To whom correspondence should be addressed. Tel: + 34 963 544 993; Email: carlos.llorens@uv.es

informative database established within an evolutionary context of classification, where one piece of research delivers one conclusion that drives individuals towards another goal. The most captivating aspect of this project is that a share of our efforts are dedicated to the interpretation of analyses, paying particular attention to non-redundant elements displaying a certain degree of distance and investigating how they can be collectively aligned or related, in terms of protein domain architecture, with other lineages and elements. Because of the impressive molecular diversity of viruses and MGEs, the GyDB is a long-term project that has been arranged in a database in continuous progression, and must be achieved in stages. The current database stage and scope is retroviruses and retrotransposons with long terminal repeats (LTR retroelements) and their relatives. Following the outline of the earlier release (the study of *Ty3/Gypsy* and *Retroviridae* LTR retroelements), this article presents the GyDB update based on the phylogenetic evaluation of the most representative LTR retroelement families and the plant caulimoviruses. This update, called GyDB 2.0, is available at http://gydb.org and includes sequence phylogenetic classification in addition to significant bioinformatic improvements. In particular, the new infrastructure implements a wiki management system constructed with the aim of promoting a world-wide community of researchers collaborating in the analysis and classification of MGEs and viruses inhabiting (or circulating in) living organisms.

THE UPDATE: NEW FEATURES

GyDB 2.0 consists of 1234 web pages addressing the phylogenetic study of *Ty3/Gypsy*, *Retroviridae*, *Ty1/Copia* and *Bel/Pao* LTR retroelement. Caulimoviruses (*Caulimoviridae*) are formally plant DNA pararetroviruses, but they were considered in GyDB 2.0 owing to their relationship with LTR retroelements based on the common gag/coat and pol regions [for more details, see (7) and references therein]. Table 1 summarizes the topics addressed in this update, as well as the servers and database sections it offers. The sequences on which GyDB 2.0 is based were retrieved from GenBank (8) and the methodologies employed were the same as those described earlier in references (6,7,9). At GyDB we evaluate the phylogenetic signal of classified distinct elements and create hidden Markov model (HMMs) profiles (10) per lineage and protein domain. In addition, the project is concerned with the evolutionary relationships between MGEs and their host genomes, based on the analysis of common protein families. In this regard, GyDB 2.0 focuses on two protein superfamilies including protein products commonly encoded by LTR retroelements and their host genomes; the chromodomain superfamily (11) and clan AA of aspartic peptidases (12,13). This second release is accompanied by bibliographic data-mining from PubMed databases hosted at the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.

Table 1. GyDB 2.0 new features: topics and contents

Systems	Families	Lineages	Elements	Protein domains	Accessory proteins	LTRs
LTR retroelements	*Ty3/Gypsy*	34	96	8	1	Yes
LTR retroelements	*Ty1/Copia*	19	69	8	–	Yes
LTR retroelements	*Retroviridae*	8	50	8	41	Yes
LTR retroelements	*Bel/Pao*	5	23	7	–	Yes
LTR retroelements[a]	*Caulimoviridae*	6	30	10	27	No
Related families	Clan AA	35	323	1	–	No
Related families	Chromodomains	2	123	1	–	No

Topics	Sections	Availability
Systematics	9	Side menu
Domains	14	Side menu
Database	8	Side menu
Servers	3	Top menu: BLAST, HMM, Literature
Wiki tools and utilities	3	Top menu

Databases	Items		Sections
Genomes (full-length genomes)	271	sequences	BLAST search and RefSeq DBs
LTRs (nucleotide sequences)	413	sequences	BLAST search and RefSeq DBs
Cores (protein cores sequences)	1895	sequences	BLAST search and RefSeq DBs
HMMs	314	HMM profiles	HMM search and GyDB collection
Multiple alignments	131	alignments	GyDB collection
Consensus sequences	314	MRC sequences	GyDB collection
Phylogenetic trees	70	trees	Phylogenies
Clan AA ancestral reconstruction	70	alignments	CAARD database
Literature	100797	references	Literature server

[a]We included caulimoviruses in the second release in view of their relationship with LTR retroelements based on the common gag/coat and pol region.

nih.gov/) to document up to date information regarding the distinct classified elements.

DATABASE ORGANIZATION

GyDB 2.0 is deployed over a Linux-MySQL-Apache-PHP (LAMP) stack, with additional Ajax programming to minimize server responses to client browsers. The design is similar to that of the previous release but implements various changes on the web interface. As shown in Figure 1, the database organization is founded upon two major menus—a top menu and a side menu. The top menu allows access to the three servers:

(i) BLAST server; implements a BLAST search powered by the NCBI BLAST package (14), allowing protein and DNA comparisons with the GENOMES, LTRs and CORES databases. These databases collect the full-length genomes, the LTR sequences and all the protein sequences on which the second release is based, respectively.
(ii) HMM server; implements HMMER3 package (http://hmmer.janelia.org) and allows protein

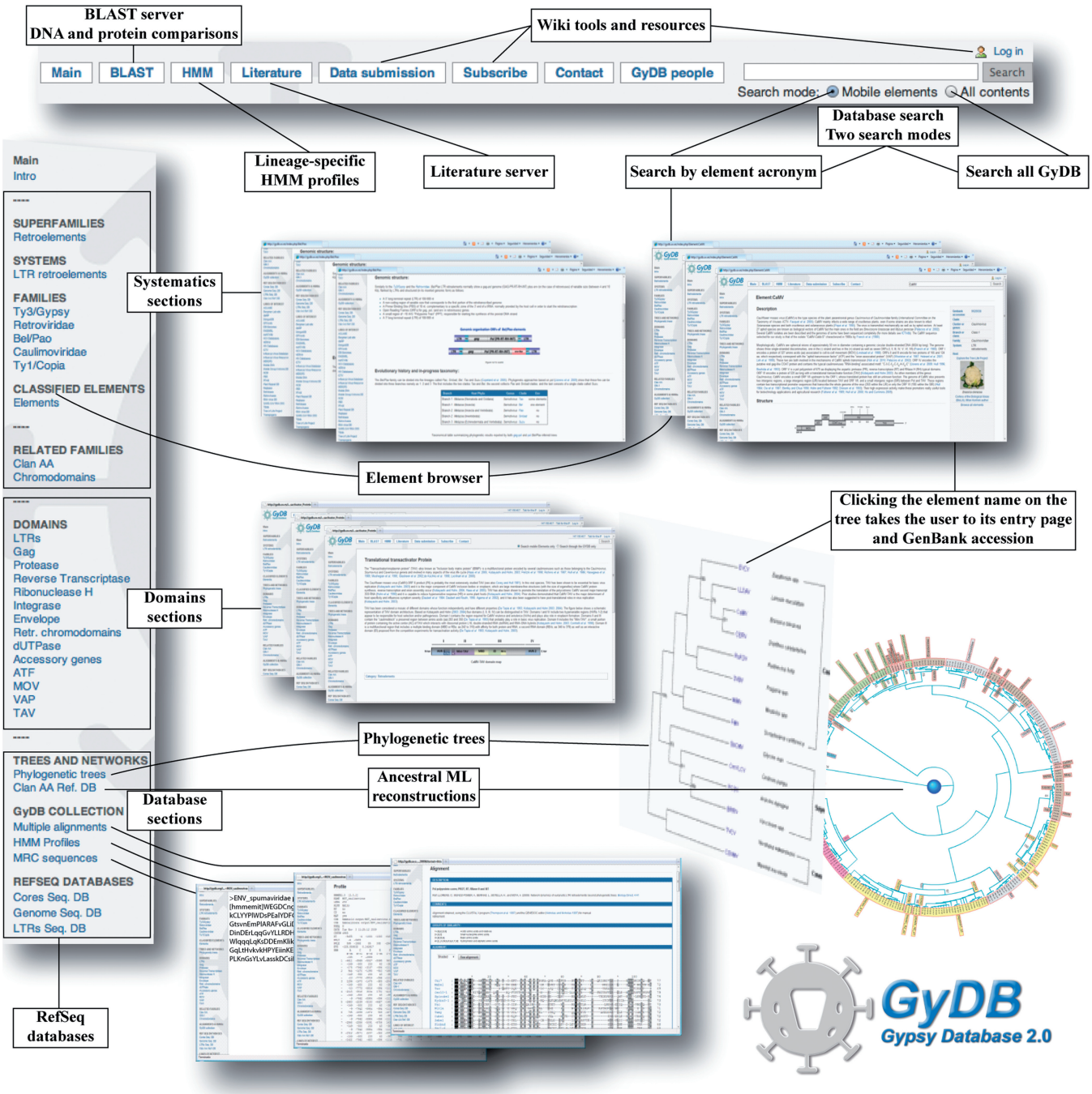

Figure 1. GyDB 2.0 organization and implementation.

comparisons against a database of protein domain lineage-specific HMM profiles created based on the update. This server provides additional comparisons between HMM profiles and the aforementioned CORES database.

(iii) LITERATURE server; allows users to search bibliography of interest in the topic.

An additional new tool in GyDB 2.0 is its wiki, powered by the MediaWiki content management system (http://www.mediawiki.org/). This tool has been implemented to allow other users participate in the project by editing or creating topics. Accession to this wiki is free but it requires a subscription (registration). The rationale behind this choice is that edits are registered by date and author in order to credit contributions, and secondly, we have programmed a revision mechanism to review all changes constructively before making them public. The top menu includes three sections to log in and manage the distinct wiki resources. Finally, to the right of the top menu, GyDB 2.0 includes a text field to search the whole project under two modes (detailed in Figure 1). The side menu divides the distinct GyDB sections into three major demarcations (emphasized with boxes in Figure 1). The first collects sections associated with the systematics applied at GyDB. The second implements information concerning the domains typically observed in the genomic structure of the elements we classify. The third demarcation offers free access to distinct databases, which are organized into three sections:

(i) Trees and Networks; consists of the collection of inferred phylogenetic trees based on distinct protein domains encoded by the classified elements, or based on their concatenation (when they are parts of polyproteins). Remarkably, inferred pol polyprotein phylogenies based on the concatenation of the protease, reverse transcriptase, RNaseH and integrase domains, are the major criterion for assigning phylogenetic levels at GyDB 2.0 [results introduced in (7)]. Phylogenetic trees provide links to the corresponding element page at GyDB 2.0. By clicking any element name in any tree an entry assigned to this element is opened. These tree image maps were created using Phylograph 1.0 (15). This section includes the clan AA reference database (CAARD) of ancestral maximum likelihood (ML) reconstructions (13) that has been implemented and maintained at GyDB.

(ii) GyDB collection (16) or the repository of multiple alignments, HMMs, and majority rule consensus (MRC) sequences offered at GyDB 2.0. When a deposited alignment, profile or MRC sequence is associated with a journal publication, its entry in the collection includes citation information.

(iii) REF SEQ DATABASES or the repository for downloading the databases (GENOMES, CORES and LTRs) implemented in the BLAST server.

Finally, a variety of links to other database initiatives relevant to the topic are included in the side menu.

FUTURE PERSPECTIVES

Sequencing projects constantly deliver new types of MGEs [for example (17–22)]; hence the classification of non-redundant elements based on their phylogenetic signal is an open issue at GyDB, and results in the preparation of new sections. For example, we are committed to improving the understanding of the diversity and evolutionary dynamics of MGEs in eukaryotic and prokaryotic organisms. In this regard of eukaryotic LTR retroelements (the current database scope), the sequence repertoire at GyDB with representative elements retrieved from recently sequenced marine secondary endosymbionts including the brown alga *Ectocarpus siliculosus* (heterokont) and the coccolithophore *Emiliania huxleyi* (haptophyte) will be implemented. In terms of other research topics in preparation, one concerns the construction of a server devoted to the study of the complete set of MGEs and repeats (the mobilome) of biological genomes. This server will be introduced with two forthcoming publications focusing on the LTR retroelements and their related transposases of the pea aphid *Acyrthosiphon pisum* genome [see (23)]. At the technical level, we are exploring the application of formal grammars and machine learning algorithms to automate, as far as possible, the management and classification of the sequence data. We are also committed to developing solutions for other non-trivial difficulties that arise with the growing size of the databases. Viruses and MGEs usually show different rates of evolution and high variability depending on the evaluated protein or region. Therefore, we aim to implement more than one method of phylogenetic reconstruction to offer the user different perspectives based on different methods (or the opportunity to upload updated phylogenies via the wiki). On the other hand, the traditional view of the origin and evolution of biological systems is that they are usually monophyletic, but such an assumption has been challenged by increasing evidence suggesting that natural evolution can frequently proceed by gradual and vertical means, in addition to distinct modular, saltatory and reticulate events (24–36). In this respect, we are investigating appropriate protocols to combine phylogenetic inference with new tendencies in network biology [see also (7)].

ACKNOWLEDGEMENTS

We thank all the colleagues detailed in the list available at (http://gydb.org/index.php/Acknowledgments) for their support in contributing images of biological host organisms. We are also grateful to Senior NAR Editor Dr Michael Galperin and to the two anonymous reviewers for their constructive comments in improving this article. Finally we also thank Denys Wheatley and Angela Panther from Biomedes for copyediting of this article.

FUNDING

Centro de Desarrollo Tecnológico Industrial (CDTI) (grant IDI-20100007, partial); Empresa Nacional de Innovación, S.A (ENISA) (17092008, partial); IMPIVA

(IMIDTA/2009/118 and IMDTA/2010/740, partial); European Regional Development Fund (ERDF); Ministerio de Ciencia e Innovación (MICINN) (Torres-Quevedo grants PTQ-09-01-00020, PTQ-09-01-00670 and PTQ-10-03552, partial). Funding for open access charge: University of Valencia.

Conflict of interest statement. None declared.

REFERENCES

1. Hurst,G.D.D. and Schilthuizen,M. (1998) Selfish genetic elements and speciation. *Heredity*, **80**, 2–8.
2. Volff,J.N. and Brosius,J. (2007) Modern genomes with retro-look: retrotransposed elements, retroposition and the origin of new genes. *Genome Dyn.*, **3**, 175–190.
3. Fauquet,C.M., Mayo,M.A., Desselberger,U. and Ball,L.A. (2005) *Virus Taxonomy, VIIIth Report of the ICTV*. Elsevier/Academic Press, London.
4. Jurka,J., Kapitonov,V.V., Pavlicek,A., Klonowski,P., Kohany,O. and Walichiewicz,J. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.*, **110**, 462–467.
5. Leplae,R., Hebrant,A., Wodak,S.J. and Toussaint,A. (2004) ACLAME: a CLAssification of Mobile genetic Elements. *Nucleic Acids Res.*, **32**, D45–D49.
6. Llorens,C., Futami,R., Bezemer,D. and Moya,A. (2008) The Gypsy Database (GyDB) of mobile genetic elements. *Nucleic Acids Res.*, **36**, 38–46.
7. Llorens,C., Munoz-Pomer,A., Bernad,L., Botella,H. and Moya,A. (2009) Network dynamics of eukaryotic LTR retroelements beyond phylogenetic trees. *Biol. Direct.*, **4**, 41.
8. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2009) GenBank. *Nucleic Acids Res.*, **37**, D26–D31.
9. Llorens,C., Fares,M.A. and Moya,A. (2008) Relationships of Gag–pol diversity between Ty3/Gypsy and Retroviridae LTR retroelements and the three kings hypothesis. *BMC Evol. Biol.*, **8**, 276.
10. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
11. 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.
12. Rawlings,N.D., Barrett,A.J. and Bateman,A. (2010) MEROPS: the peptidase database. *Nucleic Acids Res.*, **38**, D227–D233.
13. Llorens,C., Futami,R., Renaud,G. and Moya,A. (2009) Bioinformatic flowchart and database to investigate the origins and diversity of Clan AA peptidases. *Biol. Direct.*, **4**, 3.
14. 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.
15. Llorens,C., Futami,R., Vicente-Ripolles,M. and Moya,A. (2008) Phylograph: a multifunction Java editor for handling phylogenetic trees. *Biotechvana Bioinformatics*, Biotechvana, Valencia, SOFT: Phylograph.
16. Llorens,C., Muñoz-Pomer,A., Futami,R. and Moya,A. (2009) The GyDB Collection of Viral and Mobile Genetic Element Models. *Biotechvana Bioinformatics*, Biotechvana, Valencia, CR: GyDB Collection.
17. Piskurek,O., Nishihara,H. and Okada,N. (2008) The evolution of two partner LINE/SINE families and a full-length chromodomain-containing Ty3/Gypsy LTR element in the first reptilian genome of Anolis carolinensis. *Gene*, **441**, 111–118.
18. Novikova,O., Mayorov,V., Smyshlyaev,G., Fursov,M., Adkison,L., Pisarenko,O. and Blinov,A. (2008) Novel clades of chromodomain-containing Gypsy LTR retrotransposons from mosses (Bryophyta). *Plant J.*, **56**, 562–574.
19. Bae,Y.A., Ahn,J.S., Kim,S.H., Rhyu,M.G., Kong,Y. and Cho,S.Y. (2008) PwRn1, a novel Ty3/gypsy-like retrotransposon of Paragonimus westermani: molecular characters and its differentially preserved mobile potential according to host chromosomal polyploidy. *BMC. Genomics*, **9**, 482.
20. Gao,D., Gill,N., Kim,H.R., Walling,J.G., Zhang,W., Fan,C., Yu,Y., Ma,J., SanMiguel,P., Jiang,N. *et al.* (2009) A lineage-specific centromere retrotransposon in Oryza brachyantha. *Plant J.*, **60**, 820–831.
21. Gottlieb,A.M. and Poggio,L. (2010) Genomic screening in dioecious ''yerba mate'' tree (Ilex paraguariensis A. St. Hill., Aquifoliaceae) through representational difference analysis. *Genetica*, **138**, 567–578.
22. Maumus,F., Allen,A.E., Mhiri,C., Hu,H., Jabbari,K., Vardi,A., Grandbastien,M.A. and Bowler,C. (2009) Potential impact of stress activated retrotransposons on genome evolution in a marine diatom. *BMC Genomics*, **10**, 624.
23. The International Aphid Genomics Consortium. (2010) Genome sequence of the pea aphid Acyrthosiphon pisum. *PLoS Biol.*, **8**, e1000313.
24. 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.
25. Lerat,E., Brunet,F., Bazin,C. and Capy,P. (1999) Is the evolution of transposable elements modular? *Genetica*, **107**, 15–25.
26. 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.
27. 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.
28. Malik,H.S. and Eickbush,T.H. (2001) Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. *Genome Res.*, **11**, 1187–1197.
29. Marco,A. and Marin,I. (2008) How Athila retrotransposons survive in the Arabidopsis genome. *BMC. Genomics*, **9**, 219.
30. Rambaut,A., Posada,D., Crandall,K.A. and Holmes,E.C. (2004) The causes and consequences of HIV evolution. *Nat. Rev. Genet.*, **5**, 52–61.
31. Flavell,A.J. (1999) Long terminal repeat retrotransposons jump between species. *Proc. Natl Acad. Sci. USA*, **96**, 12211–12212.
32. Jordan,I.K., Matyunina,L.V. and McDonald,J.F. (1999) Evidence for the recent horizontal transfer of long terminal repeat retrotransposon. *Proc. Natl Acad. Sci. USA*, **96**, 12621–12625.
33. Bousalem,M., Douzery,E.J. and Seal,S.E. (2008) Taxonomy, molecular phylogeny and evolution of plant reverse transcribing viruses (family Caulimoviridae) inferred from full-length genome and reverse transcriptase sequences. *Arch. Virol.*, **153**, 1085–1102.
34. Koonin,E.V., Mushegian,A.R., Ryabov,E.V. and Dolja,V.V. (1991) Diverse groups of plant RNA and DNA viruses share related movement proteins that may possess chaperone-like activity. *J. Gen. Virol.*, **72(Pt 12)**, 2895–2903.
35. Llorens,J.V., Clark,J.B., Martinez-Garay,I., Soriano,S., deFrutos,R. and Martinez-Sebastian,M.J. (2008) Gypsy endogenous retrovirus maintains potential infectivity in several species of Drosophilids. *BMC Evol. Biol.*, **8**, 302.
36. de Setta,N., Van Sluys,M.A., Capy,P. and Carareto,C.M. (2009) Multiple invasions of Gypsy and Micropia retroelements in genus Zaprionus and melanogaster subgroup of the genus Drosophila. *BMC Evol. Biol.*, **9**, 279.