Mobile Genetic Elements 1:2, 97-102; July/August 2011; © 2011 Landes Bioscience

GyDB mobilomics

LTR retroelements and integrase-related transposons of the pea aphid *Acyrthosiphon pisum* genome

Guillermo P. Bernet,[1,*] Alfonso Muñoz-Pomer,[1,2] Laura Domínguez-Escribá,[1] Laura Covelli,[1] Lucía Bernad,[1,†] Sukanya Ramasamy,[3] Ricardo Futami,[1] Jose M. Sempere,[2] Andrés Moya[4] and Carlos Llorens[1]

[1]Biotechvana; Parc Cientific de la Universitat de València; [2]Departamento de Sistemas Informáticos y Computación (DSIC); Universitat Politècnica de València; Valencia, Spain; [3]Department of Mathematical Statistics; Chalmers University of Technology; Göteborg, Sweden; [4]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); Valencia, Spain

[†]Current address: Laboratory of Plant Molecular Biology; Rockefeller University; New York, NY USA

Key words: mobilome, *Ty3/Gypsy*, *Bel/Pao*, Ginger1, Ginger2, CIN1

Abbreviations: GyDB, gypsy database; MGE, mobile genetic element; LTR, long terminal repeat; IAGC, international aphid genomic consortium; RT, reverse transcriptase; AP, protease; MAC, macrodomain; INT, integrase; TR, transposase; GIN, gypsy integrase; BIN, Bel/Pao integrase; NJ, neighbor-joining

The Gypsy Database concerning Mobile Genetic Elements (release 2.0) is a wiki-style project devoted to the phylogenetic classification of LTR retroelements and their viral and host gene relatives characterized from distinct organisms. Furthermore, GyDB 2.0 is concerned with studying mobile elements within genomes. Therefore, an in-progress repository was created for databases with annotations of mobile genetic elements from particular genomes. This repository is called Mobilomics and the first uploaded database contains 549 LTR retroelements and related transposases which have been annotated from the genome of the Pea aphid *Acyrthosiphon pisum*. Mobilomics is accessible from the GyDB 2.0 project using the URL: http://gydb.org/index.php/Mobilomics.

Results and Conclusions

The Gypsy Database (GyDB) is a wiki-style database launched with the aim of classifying non-redundant viruses and mobile genetic elements (MGEs) on the basis of their phylogenetic profile. Owing to the diversity of these genetic agents, the GyDB is a long term project. In a previous article[1] the second release of this project (GyDB 2.0), which focuses on the evolutionary classification of retroviruses and retrotransposons with long terminal repeats (LTRs) and their viral and host gene relatives in distinct organisms, was introduced. GyDB 2.0 is also concerned with offering characterizations of MGEs within genomes. Therefore, a PHP-programmed repository of databases with MGE annotations classified per genome, named Mobilomics, has been created. As presented in **Figure 1**, the implementation and data organization of Mobilomics is based on a horizontal menu with access to three sections, "Annotations", "BLAST" and "Download". "Annotations" refers to a worksheet of columns and rows that can be accessed by clicking on the name of the organism addressed, with each row corresponding to an annotated sequence. Columns present information for each sequence including the gene identifier, classification, mapping, statistics and size. The information can be sorted alphabetically or numerically by clicking the column headers. Gene identifiers provide links that redirect the user to files containing the genomic and protein sequences and not-to-scale gene representations for all annotations. Below the worksheet, there are links to two sections that contain information regarding methods and discussion of the issue. The "BLAST" section provides access to a BLAST search[2] specific for each classified genome, allowing proteins and nucleotides to be compared between any user-defined query and the databases available for each genome. The "Download" section permits users to download these databases.

Mobilomics is a tool undergoing continuous progression, which now contains the genome of the Pea aphid *Acyrthosiphon pisum*[3] from which 549 LTR retroelement-like features were mapped and annotated. Gene identifiers in this collection follow the nomenclature provided by the International Aphid Genomic Consortium (IAGC) and this nomenclature was adopted to also deposit (as members of the IACG) annotations from our group into Aphidbase[4] version 1.2. The collection consists of full-length and fragmented elements. Almost all annotations correspond to LTR retroelement sequences classified as members of the

*Correspondence to: Guillermo Bernet; Email: guillermo.bernet@biotechvana.com
Submitted: 07/06/11; Accepted: 08/04/11
DOI: 10.4161/mge.1.2.17635

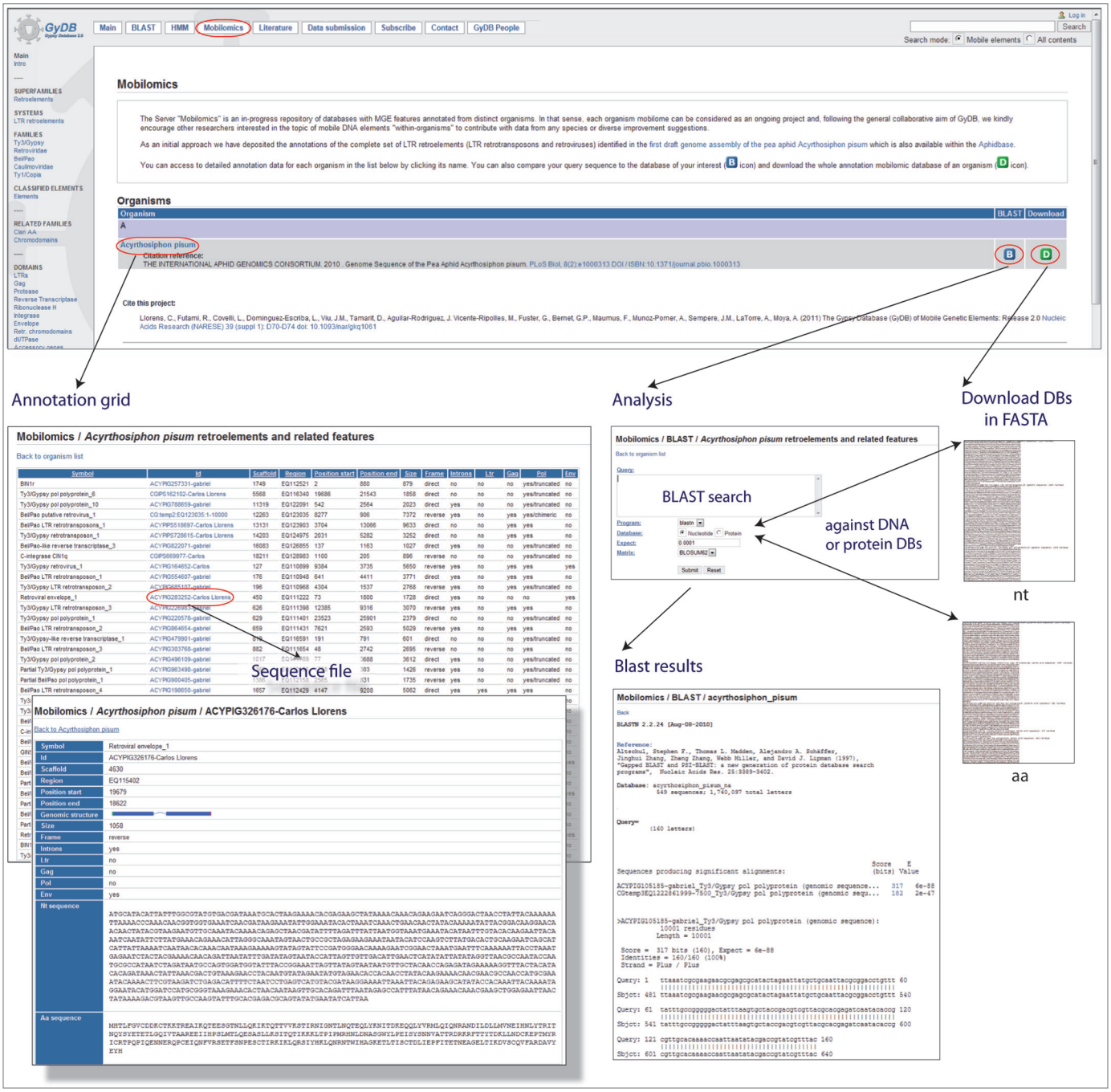


Figure 1. Mobilomics screenshot, implementation and organization.

Ty3/Gypsy and *Bel/Pao* families, and the *Ty3/Gypsy* characterization resulted in 121 full-length elements (of which eight are putative endogenous retroviruses) and 161 partial features (single gene sequences for gag, or pol, or pol-env domains) being identified. Phylogenetic reconstruction analysis based on reverse transcriptase (RT) revealed known lineages including *CsRN1*, *Errantiviridae*, *Osvaldo, Mag and Micropia/Mdg3* (**Fig. 2A**). In addition, four new clades named ATYMPO1, ATYMPO2 (Aphid *Ty3/Gypsy* element with Macro domain, Poa1p-like family), ATY3 and ATY4 (Aphid *Ty3/Gypsy* element) were identified. **Figure 2B** presents a graphical representation of these new elements. Remarkably, the ATYMPO1 and ATYMPO2 clades contain LTR retrotransposons and retroviruses with typical LTR retroelement regions but differing from other *Ty3/Gypsy* elements as they possess an additional pol-like domain between the protease (AP) and RT. BLASTp[2] searches performed using the NCBI protein databases revealed similarity (9.06e-17) between this domain and the Macro (MAC) domain of the Poa1p family (classified as cd02901 in the NCBI Conserved Domain Databases[5]). MAC is

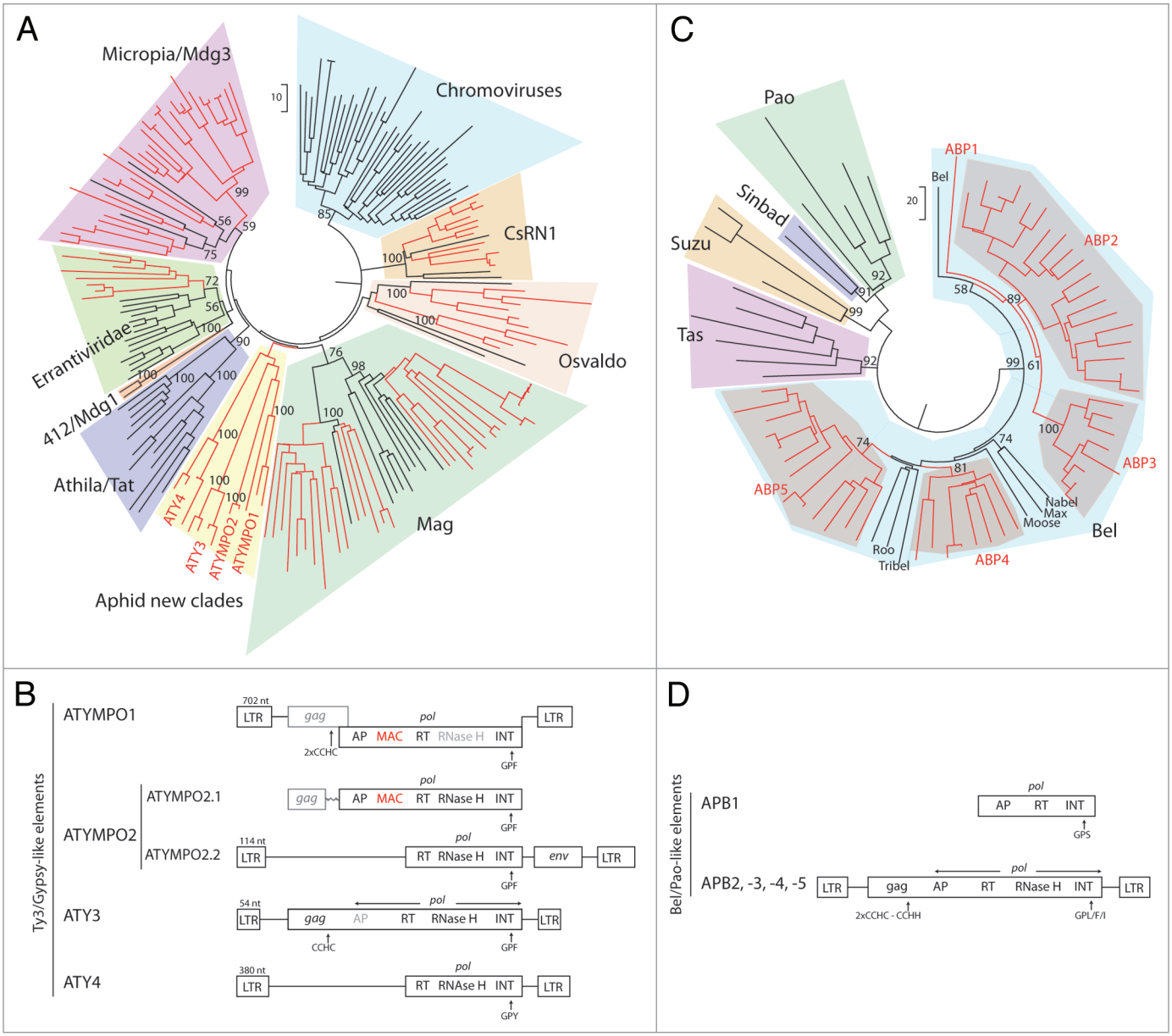


Figure 2. (A) Inferred phylogenetic tree of *Ty3/Gypsy* LTR retroelements based on an RT alignment using Mega[20] and the Neighbor-joining method of phylogenetic reconstruction. Alignment methods were identical to those described in reference 9. The alignment is deposited in the GyDB collection with the following URL: http://gydb.org/index.php/Collection_alignments. Black edges refer to known *Ty3/Gypsy* elements used as a phylogenetic reference for each clade; red edges correspond to the evaluated *A. pisum* sequences. Bootstrap values higher than 50% are indicated at the intersection of relevant branches. (B) New *Ty3/Gypsy* clades. Open boxes define the presence of relevant structural features and domains (structures not to scale). (C) Phylogenetic tree of *Bel/Pao* LTR retroelements based on the RT using identical methods to those described above. The alignment used to infer this tree has been deposited in the GyDB collection with the URL: http://gydb.org/index.php/Collection_alignments. Black edges refer to reference *Bel/Pao* elements and red edges correspond to *A. pisum* sequences. (D) New *Bel/Pao* clades that, with the exception of ABP1 (which is a pol remnant), contain full-length elements with typical features of *Bel/Pao* LTR retroelements.

a high-affinity ADP-ribose binding module present in eukaryotes, prokaryotes and diverse RNA viruses, such as coronaviruses and alphaviruses, which replicate in the cytoplasm of animal cells.[6] LTR retroelement-sequences with a MAC domain have only been reported in diverse *Danio rerio* retroviruses, where MAC maps between the Integrase (INT) and the ENV domains.[7] Therefore, as far as we are aware, this is the first time that insect *Ty3/Gypsy* elements have been reported to contain MAC. Regarding *Bel/Pao* LTR retroelements, 86 full-length elements (two being putative retroviruses) and 92 additional partial *Bel/Pao*-like features, were annotated. In accordance with references 8 and 9, the *Bel/Pao* family can be divided into five major lineages—*Bel, Suzu, Sinbad, Pao* and *Tas*. As demonstrated in **Figure 2C**, all *Bel/Pao* sequences that were analyzed (on the basis of RT) are related to the *Bel* lineage. However, almost all the *Bel/Pao* elements that were characterized

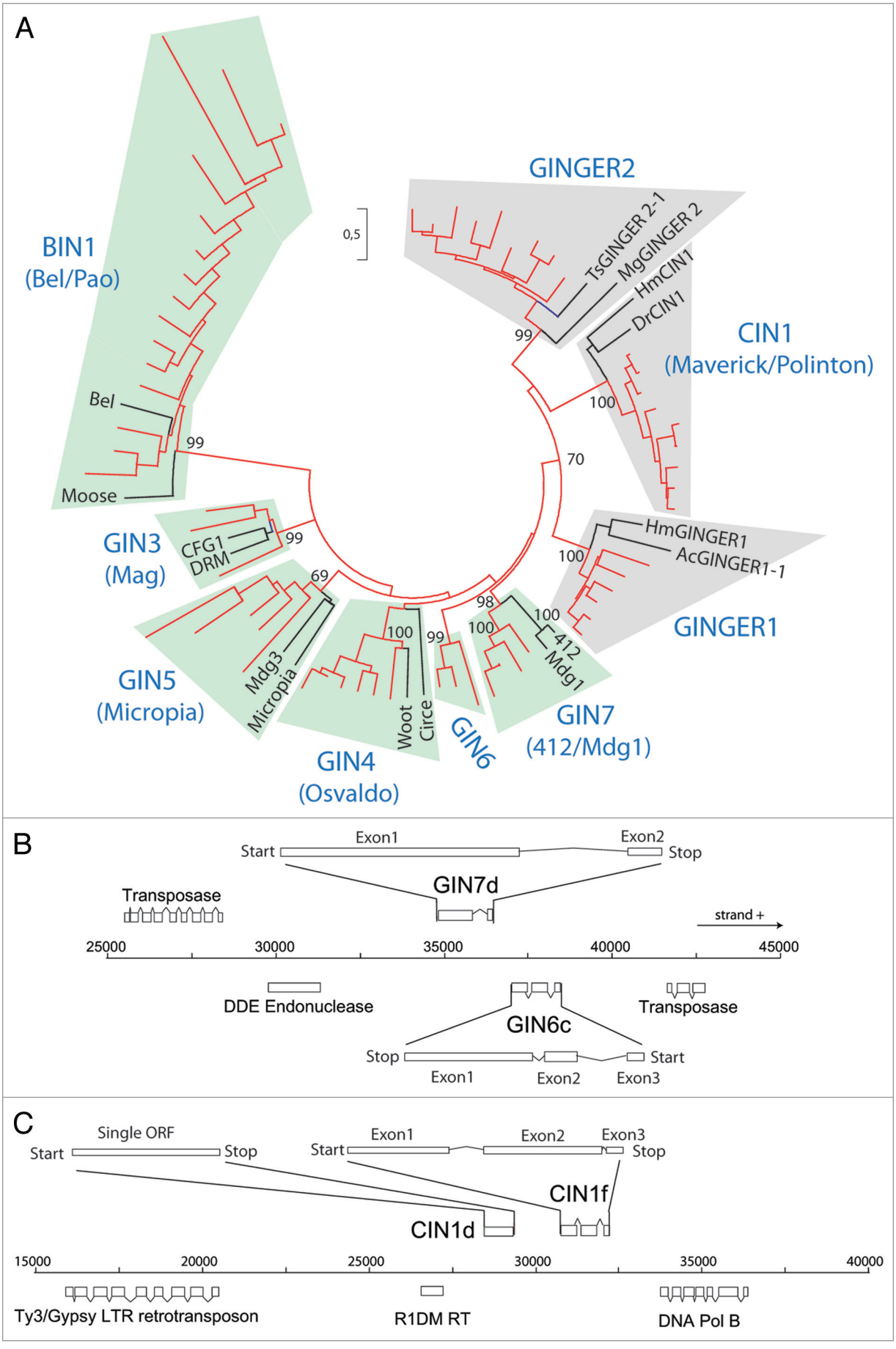
Figure 3 (See opposite page). (A) Inferred phylogenetic tree based on solo-INTs and TRs of *A. pisum* using identical methods to those described in **Figure 2**. The alignment used to infer this tree has been deposited in the GyDB collection with the URL: http://gydb.org/index.php/Collection_alignments. Red edges correspond to *A. pisum* sequences and black edges refer to reference INTs and TRs. (B) Genomic continuum representation of the scaffold 2,047 of the first aphid assembly, where GIN7d and GIN6c (nucleotides 34,807–36,407 and 38,019–36,918, respectively) map close to other MGEs not related to LTR retroelements. (C) Genomic continuum representation of the scaffold 6092, where the *Maverick/Polinton*-like TRs CIN1d and CIN1f map (nucleotides 28,451–29,308 and 30,520–31,959 respectively) map close to other features.

were new sequences that could be divided into five clades that were named ABP1, -2, -3, -4 and -5 (Aphid *Bel/Pao* element). Graphical depictions are presented in **Figure 2D**. These new clades suggest that the diversity of the *Bel/Pao* family in Metazoans is greater than previously thought (as argued by de la Chaux and Wagner in a recent publication in ref. 10). It was interesting that although partial sequences related to the *Ty1/Copia* family were annotated, no full-length *Ty1/Copia* elements were detected in the screenings. These sequences are likely to represent reminiscences of an ancient *Ty1/Copia* colonization but it could be the case that further releases of the *A. pisum* genome (or other aphid relatives) will reveal the presence of full-length *Ty1/Copia* elements.

The aphid collection includes multiple distinct single gene features that are likely to be of interest to researchers investigating the contribution of MGEs to the complexity of their host genomes. A putative case study, based on a pool of 60 sequences, has been selected and this contains a variety of LTR retroelement Solo-INTs, diverse cut-and-paste DNA transposons of the *Ginger1* and *-2* families,[11] and other sequences identified as the chromodomain-carrying transposases (TRs) typically encoded by the *Maverick/Polinton* transposons.[12,13] The latter TRs were classified using the term CIN1 (chromodomain-INTs type 1) as no apparent full-length *Maverick/Polinton* element was detected in the first *A. pisum* release. CIN1 TRs are therefore putative remnants derived from an ancient *Maverick* colonization. Interestingly, it is now known that an evolutionary link exists among Ginger-like and CIN1 TRs and LTR retroelement INTs, but this is also the case for other DNA transposons and INT-like host genes that are probably derived from the domestication of MGEs (reviewed in ref. 11 and 14–18). All of these are DDE INTs and TRs that, on the basis of INT-like structural or sequence similarities, can be considered as representative members of the Retroviral Integrase Superfamily[19] of nucleic acid-processing enzymes involved in selfish evolution, replication and repair of DNA, recombination and gene fusion, RNA-mediated gene silencing and oncogenesis. **Figure 3A** illustrates the dynamics of these enzymes in the aphid genome through an inferred phylogenetic reconstruction analysis based on the Ginger1, Ginger2 and CIN1 TRs and Solo-INTs of *A. pisum.* Note that the term "Solo-INT" is used to describe single gene annotations of LTR retroelement-like INTs that do not map close to any other typical LTR retroelement domain in the aphid genome. In other words, these INTs are sequences excised from their LTR retroelement carriers but preserving a recognizable phylogenetic signal of their retrotransposable past. Eight clades of LTR retroelement Solo-INTs were characterized; seven of these (GIN3, -4, -5, -6 and -7) were Ty3/Gypsy INTs related to the *Mag, Osvaldo, Micropia/Mdg3* and *412/Mdg1* clades of LTR retroelements, respectively. To annotate these sequences, the term GIN-like (Gypsy INT type) was employed, following the nomenclature used in previous issues when describing the GIN1 and -2 genes in other taxa.[11,15,17] The other clade contains Solo-INTs related to *Bel/Pao* LTR retroelements and was called BIN1 (Bel INT type 1) using a similar criterion. At the phylogenetic level, **Figure 3A** demonstrates that CIN1 and Ginger2 aphid-like TRs are close to one another and that Ginger1 TRs represent an intermediate state between the two pools of TRs and the Ty3/Gypsy-like Solo-INTs. Interestingly, *Ginger1* and *Ginger2* elements are considered to be transposons because of the presence of recognizable terminal inverted repeats and flanking target site duplication sequences.[11,17] With the exception of two loci belonging to the GIN5 clade (which were flanked by small LTRs), no apparent LTRs, inverted repeats and/or target site duplications were identified in any CIN1, GIN-like or BIN1 loci. However, almost all the annotated GIN-like, BIN1 and CIN1 sequences were apparently functional loci (few possessed stop codons). Furthermore, similar to Ginger-like elements, almost all the annotated GIN-like, BIN1 and CIN1 sequences represent intron-exon structured genes that spread in the aphid genome as interdispersed single features or constituted small tandems that normally mapped close to other transposons. **Figure 3B and C** provide two examples concerning two GIN-like and CIN1-like annotations. It is difficult to elucidate whether these loci play a role in the biology of the aphid but the aphid genome is particularly rich in such loci. Therefore, it is possible that the aforementioned pools of TRs and INTs (in addition to the aphid genome) will be valuable for researching the mechanisms employed by a host genome to recycle junk DNA into new pools of transposons and/or host genes. This material is available for researchers within the *A. pisum* collection of Mobilomics, together with all other LTR retroelement annotations. Mobilomics is a tool undergoing continuous progression and further updates of the *A. pisum* collection, based on the characterization of other MGEs and two additional issues, are planned. One of these issues concerns the characterization of the mobilome of the aphid *Cinara cedri* genome, a sequencing project conducted by the IAGC. The other issue concerns revision of what is currently known about the human mobilome. Mobilomics is accessible at: http://gydb.org/index.php/Mobilomics.

Acknowledgments

We thank Denys Wheatley and Angela Panther from Biomedes for copyediting of this article, which was partially supported by Grants IDI-20100007 from CDTI (Centro de Desarrollo Tecnológico Industrial), IMIDTA/2009/118 from IMPIVA and ERDF (European Regional


A
BIN1
(Bel/Pao)
Bel
Moose
99
GIN3
(Mag)
CFG1
DRM
99
GIN5
(Micropia)
Mdg3
Micropia
69
GIN4
(Osvaldo)
Woot
Circe
100
99
GIN6
98
100
GIN7
(412/Mdg1)
412
Mdg1
100
GINGER1
HmGINGER1
AcGINGER1-1
100
70
CIN1
(Maverick/Polinton)
100
GINGER2
TsGINGER2-1
MgGINGER2
HmCIN1
DrCIN1
99
0,5
B
Start
Exon1
Exon2
Stop
GIN7d
Transposase
strand +
25000
30000
35000
40000
45000
DDE Endonuclease
GIN6c
Transposase
Stop
Start
Exon1
Exon2
Exon3
C
Single ORF
Start
Stop
Start
Exon1
Exon2
Exon3
Stop
CIN1f
CIN1d
15000
20000
25000
30000
35000
40000
Ty3/Gypsy LTR retrotransposon
R1DM RT
DNA Pol B

Development Fund), Torres-Quevedo Grants PTQ-09-01-00020, and PTQ-10-03552 from MICINN (Ministerio de Ciencia e Innovación) and grant Prometeo/2009/092 (Conselleria d'Educació, Generalitat Valenciana, Spain).

References

1. Llorens C, Futami R, Covelli L, Dominguez-Escriba L, Viu JM, Tamarit D, et al. The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. Nucleic Acids Res 2011; 39:70-4.
2. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997; 25:3389-402.
3. The International Aphid Genomics Consortium. Genome sequence of the pea aphid *Acyrthosiphon pisum*. PLoS Biol 2010; 8:e1000313.
4. Legeai F, Shigenobu S, Gauthier JP, Colbourne J, Rispe C, Collin O, et al. AphidBase: a centralized bioinformatic resource for annotation of the pea aphid genome. Insect Mol Biol 2010; 19:5-12.
5. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, et al. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res 2011; 39:225-9.
6. Neuvonen M, Ahola T. Differential activities of cellular and viral macro domain proteins in binding of ADP-ribose metabolites. J Mol Biol 2009; 385:212-25.
7. Basta HA, Cleveland SB, Clinton RA, Dimitrov AG, McClure MA. Evolution of teleost fish retroviruses: characterization of new retroviruses with cellular genes. J Virol 2009; 83:10152-62.
8. Copeland CS, Mann VH, Morales ME, Kalinna BH, Brindley PJ. The Sinbad retrotransposon from the genome of the human blood fluke, *Schistosoma mansoni*, and the distribution of related Pao-like elements. BMC Evol Biol 2005; 5:20.
9. Llorens C, Munoz-Pomer A, Bernad L, Botella H, Moya A. Network dynamics of eukaryotic LTR retroelements beyond phylogenetic trees. Biol Direct 2009; 4:41.
10. de la Chaux N, Wagner A. BEL/Pao retrotransposons in metazoan genomes. BMC Evol Biol 2011; 11:154.
11. Bao W, Kapitonov VV, Jurka J. Ginger DNA transposons in eukaryotes and their evolutionary relationships with long terminal repeat retrotransposons. Mob DNA 2010; 1:3.
12. Pritham EJ, Putliwala T, Feschotte C. Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. Gene 2007; 390:3-17.
13. Kapitonov VV, Jurka J. Self-synthesizing DNA transposons in eukaryotes. Proc Natl Acad Sci USA 2006; 103:4540-5.
14. Capy P, Langin T, Higuet D, Maurer P, Bazin C. Do the integrases of LTR-retrotransposons and class II element transposases have a common ancestor? Genetica 1997; 100:63-72.
15. Llorens C, Marin I. A mammalian gene evolved from the integrase domain of an LTR retrotransposon. Mol Biol Evol 2001; 18:1597-600.
16. Wells DJ. Tdd-4, a DNA transposon of Dictyostelium that encodes proteins similar to LTR retroelement integrases. Nucleic Acids Res 1999; 27:2408-15.
17. Marin I. GIN transposons: Genetic elements linking retrotransposons and genes. Mol Biol Evol 2010; 27:1903-11.
18. Gao X, Voytas DF. A eukaryotic gene family related to retroelement integrases. Trends Genet 2005; 21:133-7.
19. Nowotny M. Retroviral integrase superfamily: the structural perspective. EMBO Rep 2009; 10:144-51.
20. Kumar S, Nei M, Dudley J, Tamura K. MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief Bioinform 2008; 9:299-306.