Supporting Information

Chang *et al.* 10.1073/pnas.0803860105

SI Materials and Methods

CDD Profile Database. The CDD (Conserved Domain Database) was used as the knowledge base for this study. The entire database (released in May 2008) was downloaded, including 24,280 PSSMs (position specific scoring matrices) and its CDD consensus sequences at the CDD website of http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml. We installed this database into the local LION-XC server that is administrated by the High Performance Computing Group at Penn State (see more information at the web site of http://gears.aset.psu.edu/hpc/). Also, we downloaded standalone BLAST executables package of 2.2.18 version from the NCBI BLAST ftp site (ftp://ftp.ncbi.nlm.nih.gov/blast/), which includes rps-BLAST, formatrpsdb, etc.

GDDA-BLAST. To define several generalized terms for use, it is assumed that N is the number of queries (primary amino acid sequences) of interest, and M is the number of domain profiles that were used for knowledge-based analysis of protein sequences. If the sequence length of the ith query is l_i, where i is the index number of a query, from 1 to N, the sum of the length of all of the queries is

$$l_{\text{total}} = \sum_{i=1}^{N} l_i.$$

GDDA-BLAST methodology is described below in the two separate steps: (1) signal encoding and (2) signal analysis.

Step 1. Signal Encoding. ***Query modification.*** Modification of a query sequence is achieved by inserting a sequence segment (i.e., "seed") derived from a target sequence before sequence comparison. In this study, all of the seeds were extracted from 24,280 domain profiles downloaded from CDD. The size of a seed is fixed as a residue number or the fraction of its sequence length in the consensus sequence. Two types of seeds are used; one is extracted at the N-terminal end of a particular consensus sequence, and another at the C-terminal end (i.e., "N-terminal seed" or "C-terminal seed" respectively). The resulting seeds of a profile are separately inserted between each residue position of a query sequence. Fig. 1*B*i-ii shows that the number of modified query sequences generated from a query and a profile is twice the sequence length of the query ($2 \times l_i$) because of the two types of seeds used. So, the total $2N \times l_{total}$ modified sequences per a profile are finally built with N queries for analysis

Signal collection. In GDDA-BLAST, specific signals of interest are collected during high-throughput alignment between the modified queries and the profiles by rps-BLAST. In this study, three kinds of signals encoded in the optimal alignments were computed, i.e., percentage sequence identity, percentage domain coverage, and the normalized hit number (ratio of the total number of alignments to the modified query number, scaled between 0–100) (see Fig. 1*B*iv). The percentage identity is the sequence identity in the optimal alignment found, where the identical residues included in the inserted seed are counted (Eq. 1). The percent coverage is the percentage of the length of an optimal alignment (exactly expressed as $Q_{\text{end}} - Q_{\text{start}} + 1$ in Eq. 1) to the full length of the domain profile (Eq. 1). In a given search, rps-BLAST reports optimal local alignments between modified queries and profiles with high scores. Basically, simple thresholds are adopted for filtering alignments that may be false positives. In this study, the minimum 60% domain coverage and 10% sequence identity have been chosen as the default thresholds. The optimal alignment satisfying the given thresholds is recoded as a hit. The hit number is counted with the hits found between a given query (having $2 \times l_i$ modified sequences) and a given domain profile.

$$\text{Coverage } (\%) = \frac{Q_{\text{end}} - Q_{\text{start}} + 1}{l_{\text{profile}}} \times 100 \quad [1]$$

$$\text{Identity } (\%) = \frac{N_{\text{identical}}}{l_{\text{alignment}}} \times 100 \quad [2]$$

$$H_{\text{normalized}} = \frac{N_{\text{hit}}}{l_{\text{query}} \times 2} \times 100 \quad [3]$$

where l_{profile} = The length of the consensus sequence of a profile

l_{query} = The sequence length of a query

($l_{\text{query}} \times 2$ = the total number of all the modified sequences of the query)

N_{hit} = The number of the hits from the modified query sequences given the threshold

$N_{\text{identical}}$ = The number of identical residues in the alignment

Q_{start} = The index number where an alignment starts in the modified query sequence

Q_{end} = The index number where an alignment ends in the modified query sequence

The profile that has at least one hit is called "positive" for the query, and the "negative" profile means that it has no optimal alignment satisfying the particular thresholds. During sequence comparison by rps-BLAST, the three signals are collected and finally encoded together into a working data space, which will be described in the next section.

Data space formation. The three types of signals (i.e., the normalized hit number, percentage coverage, and percentage identity) collected in the GDDA-BLAST are ultimately encoded into an array of N vectors with M dimension (see Fig. 1B-v). All hits are recorded for mean percentage coverage and mean percentage identity from the alignments and incorporated into an $N \times M$ microarray data matrix. An M-vector can be described as a "phylogenetic profile" of a given query, and each element of this vector contains a score that is computed from the signals of the query per a particular profile. Several types of scoring system were considered, and the composite (product) score, which is the multiplication of the hit ratio, the mean percentage coverage, and mean percentage identity, is used in this study. With respect to the control unmodified scores the composite (product) score is defined as the multiplication of the mean percentage coverage, and mean percentage identity. Based on such a data matrix, data analysis is performed by using a Euclidean distance metric (see Fig. 1*B*vi-vii).

Step 2. Signal Analysis. ***Phylogenies based on Euclidean distance matrices.*** For the inference of phylogenetic trees, we used Euclidian distance measurements. In detail, the Euclidean distance between two M vectors, each of which represents a particular

sequence, is calculated as expressed in Eq. 3. Through pairwise distance determination, an $N \times N$ distance matrix is produced and then used to construct a phylogenetic tree.

Euclidean distance between the phylogenetic profiles X and Y of two sequences, says $D(X, Y)$, is as follows:

$$D(X, Y) = |X - Y| = \sqrt{\sum_{i=1,M} (x_i - y_i)^2} \quad [3]$$

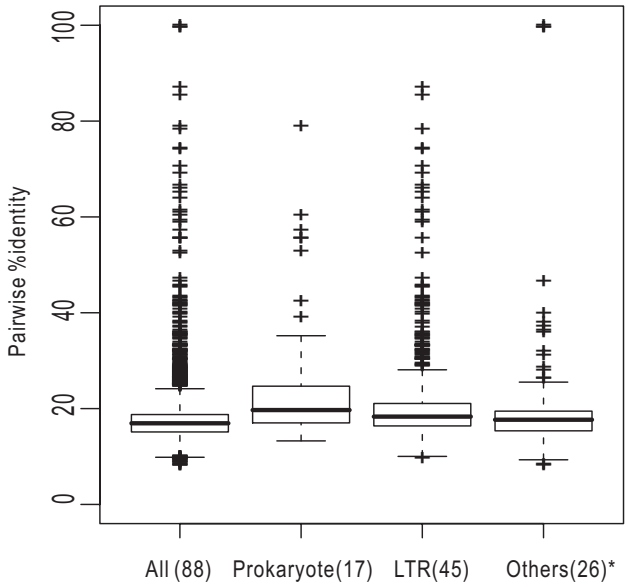
Given a Euclidean distance matrix, a phylogenetic tree of the given protein sequences is produced by using the minimum evolution (ME) method (1). The tree is estimated using MEGA4 software (http://www.megasoftware.net) (2) with the default parameter setting; the ME tree is searched using the Close-Neighbor-Interchange algorithm (3), and the Neighbor-joining algorithm (4) is used to generate the initial tree.

Statistics. ***Bootstrap test.*** For bootstrap re-sampling, 1,000 replicates were generated by M number of profile columns randomly selected from the microarray data. The same profile column was allowed to be selected more than once. The random number generator in the PHYLIP source code (http://evolution.genetics.washington.edu/phylip.html) was used to implement the code to resample GDDA-BLAST data. We also used the Fitch and Consense programs with default settings in PHYLIP 3.67 package (5) to generate minimum-evolution trees for each sample and their consensus tree by extended majority rule, respectively. Among the profiles used to measure 88 sequences: 7,768 profiles, which are about 32% of the total profiles, were "negative" in all 88 sequences. Because those profiles are not informative, we excluded these profiles columns when we generate a sample. [Total number of profile columns = 24,280, number of negative profile columns = 7,768, number of positive profile columns = 16,512 (=24,280 − 7,768)].

Jackknife test. Jackknife resampling was performed similarly to the bootstrap test. Again, 1,000 replicates were used; however, this time we sampled 80% of the original data so that the number of profile columns to select to generate each sample is 13,209 (=16,512 × 0.8). Once all of the samples were generated, we produced minimum evolution trees for all of the replicates and a consensus tree. We report the support values for each branch of our tree (Fig. 3, Fig. S2).

1. Rzhetsky A, Nei M (1993) Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Mol Biol Evol* 10:1073–1095.
2. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* 24:1596–1599.
3. Kumar S, Nei M (2000) *Molecular Evolution and Phylogenetics.* (Oxford Univ Press, Oxford, UK).
4. Saitou N, Nei M (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4:406–425.
5. Felsenstein J (1997) An alternating least squares approach to inferring phylogenies from pairwise distances. *Syst Biol* 46:101–111.

(A) Box plot of pairwise %identity of RT88 sequences



(B) Cumulative frequency distribution of pairwise %identity of RT88 sequences

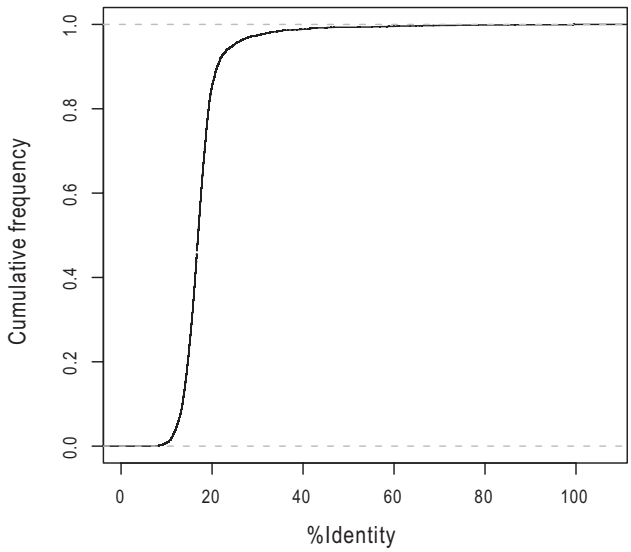


Fig. S1. Sequence similarity of the 88 retroelements. (*A*) A box plot is shown of the percentage sequence identities among the RT domain region within 88 RT sequences, and three different groupings: prokaryotic group (containing retrons, retrointrons, and diversity generating retroelements, LTR (including all LTR subgroups, DIRS1, retroviruses, caulimoviruses, and hepadnaviruses), and Other (comprising non-LTRs, retroplasmids, telomerases, and Penelope-like elements). Percentage identity was calculated based on the Needleman–Wunsch global alignment algorithm (Blosum45) of the RT domain boundary defined by GDDA-BLAST. Importantly, the majority of pairwise sequence identities among the 88 retroelements are present within the ''twilight zone'' of sequence similarity (≈15%–30% identity). (*B*) Cumulative distributions of sequence identity.

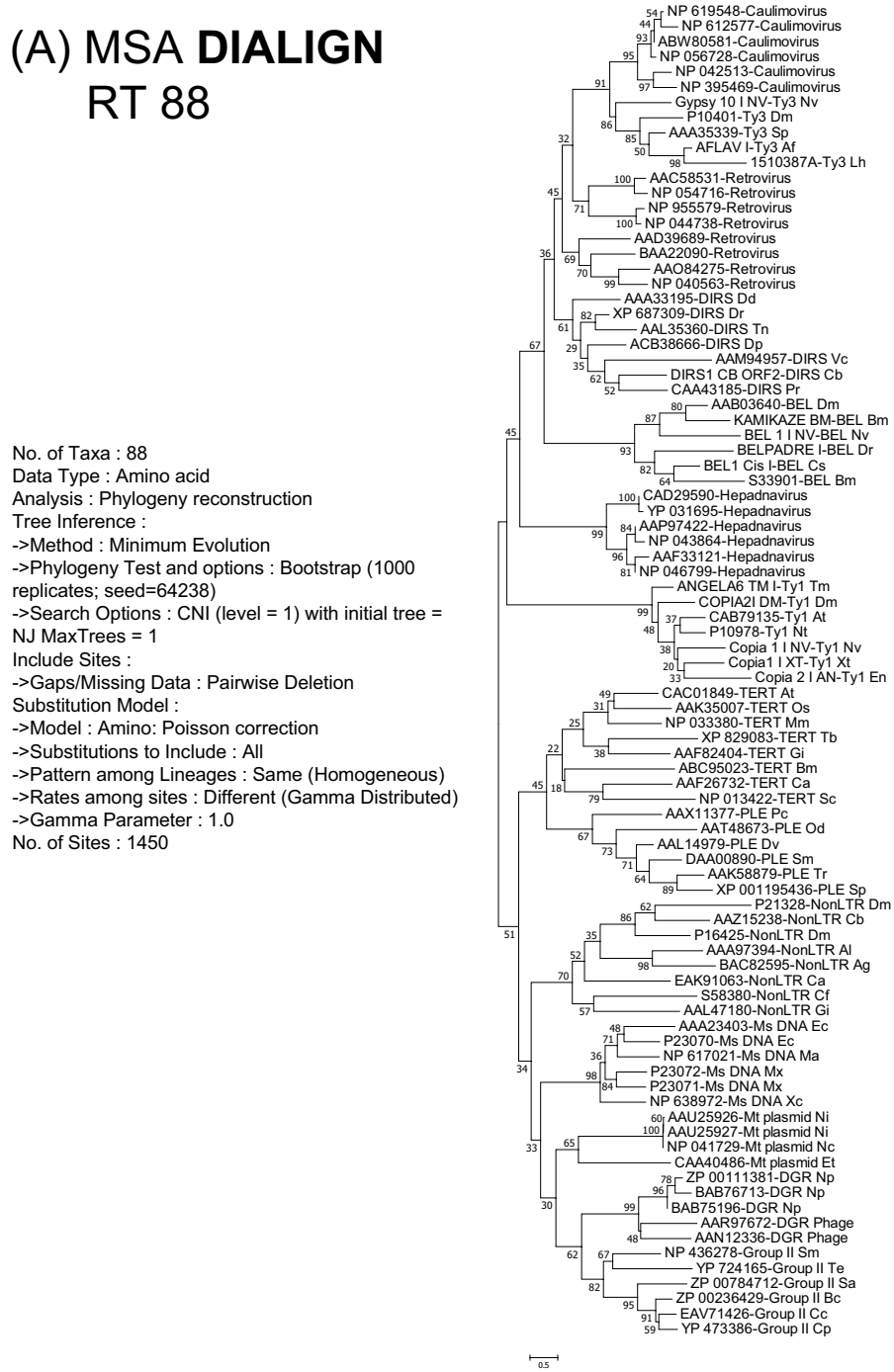


Fig. S2. Phylogenies inferred from multiple sequence alignment. Unrooted phylogenetic trees based on four different MSA methods used to align the RT domain region within 88 RT sequences. These include: (*A*) DIALIGN 2.2.1 (http://bibiserv.techfak.uni-bielefeld.de/dialign/welcome.html), (*B*) K-align (http://www.ebi.ac.uk/Tools/kalign/), (*C*) MUSCLE (http://www.ebi.ac.uk/Tools/muscle/), and (*D*) ClustalW2 (http://www.ebi.ac.uk/Tools/clustalw2/). When running each MSA method, default settings were used. Phylogenetic trees based on MSA were constructed by minimum evolution method with bootstrap support values using MEGA4. All related settings to generate each tree are presented in each figure. (*E*) An unrooted phylogenetic tree of the RT domain region within 88 RT sequences produced by the estimation of their evolutionary distances by GDDA-BLAST. This tree includes species information that was not included in main text Fig. 3. (*F*) Cartoon depicting the topology of the RT domain region within 88 RT sequences measured by GDDA-BLAST. In this case we have rooted the topology with the putative prokaryotic outgroup (red). Values at branch points denote the same topology in the MSA-based methods (e.g., a value of 100 means that all four MSA-methods show the same branching pattern). The majority of the branch points without a consensus value are due to the different placement of retroplasmids in the GDDA-BLAST derived tree.

(B) MSA **Kalign** RT 88

No. of Taxa : 88
Data Type : Amino acid
Analysis : Phylogeny reconstruction
Tree Inference :
->Method : Minimum Evolution
->Phylogeny Test and options : Bootstrap (1000 replicates; seed=64238)
->Search Options : CNI (level = 1) with initial tree = NJ MaxTrees = 1
Include Sites :
->Gaps/Missing Data : Pairwise Deletion
Substitution Model :
->Model : Amino: Poisson correction
->Substitutions to Include : All
->Pattern among Lineages : Same (Homogeneous)
->Rates among sites : Different (Gamma Distributed)
->Gamma Parameter : 1.0
No. of Sites : 908

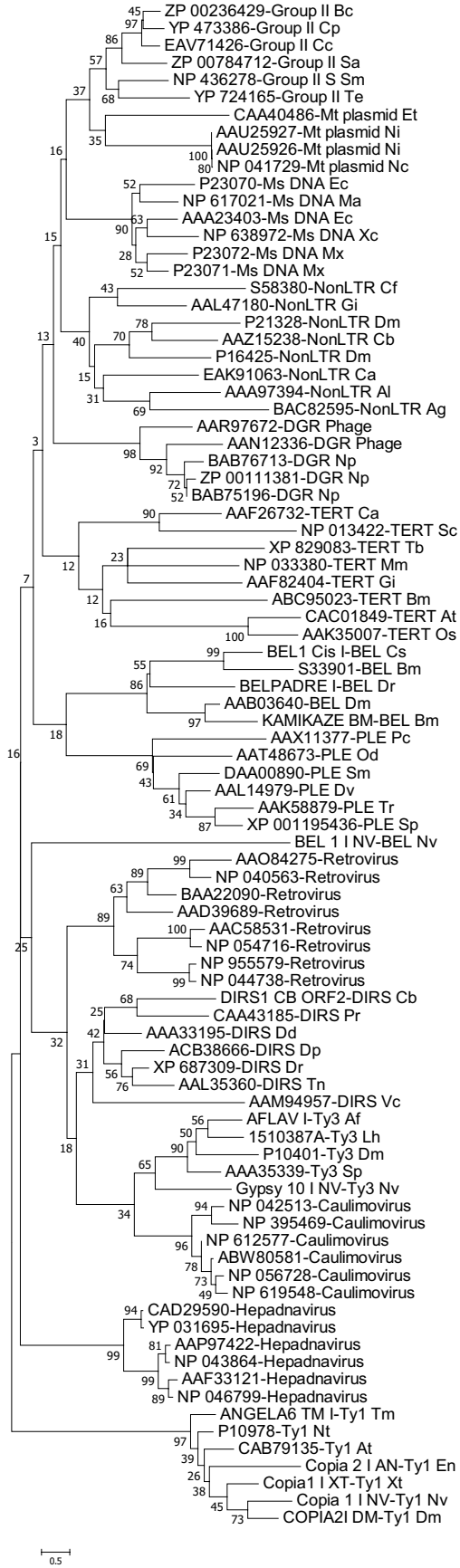


Fig. S2. (continued)

(C) MSA **MUSCLE** RT 88

No. of Taxa : 88
Data Type : Amino acid
Analysis : Phylogeny reconstruction
Tree Inference :
->Method : Minimum Evolution
->Phylogeny Test and options : Bootstrap (1000 replicates; seed=64238)
->Search Options : CNI (level = 1) with initial tree = NJ MaxTrees = 1
Include Sites :
->Gaps/Missing Data : Pairwise Deletion
Substitution Model :
->Model : Amino: Poisson correction
->Substitutions to Include : All
->Pattern among Lineages : Same (Homogeneous)
->Rates among sites : Different (Gamma Distributed)
->Gamma Parameter : 1.0
No. of Sites : 741

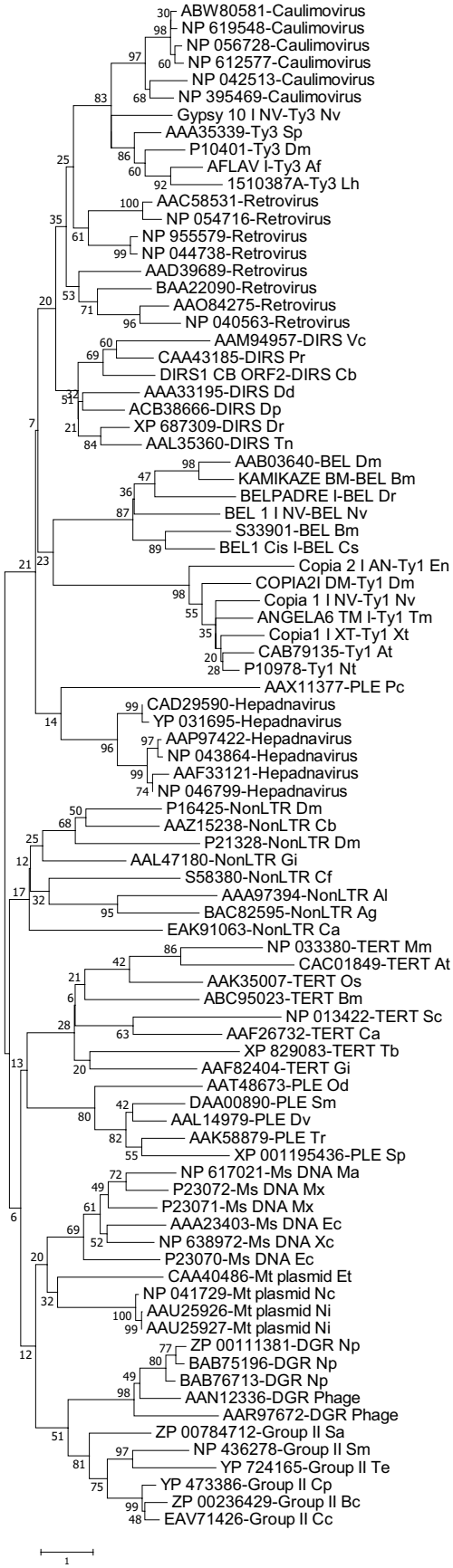


Fig. S2. (continued)

(D) MSA **ClustalW2** RT 88

No. of Taxa : 88
Data Type : Amino acid
Analysis : Phylogeny reconstruction
Tree Inference :
->Method : Minimum Evolution
->Phylogeny Test and options : Bootstrap (1000 replicates; seed=64238)
->Search Options : CNI (level = 1) with initial tree = NJ MaxTrees = 1
Include Sites :
->Gaps/Missing Data : Pairwise Deletion
Substitution Model :
->Model : Amino: Poisson correction
->Substitutions to Include : All
->Pattern among Lineages : Same (Homogeneous)
->Rates among sites : Different (Gamma Distributed)
->Gamma Parameter : 1.0
No. of Sites : 655

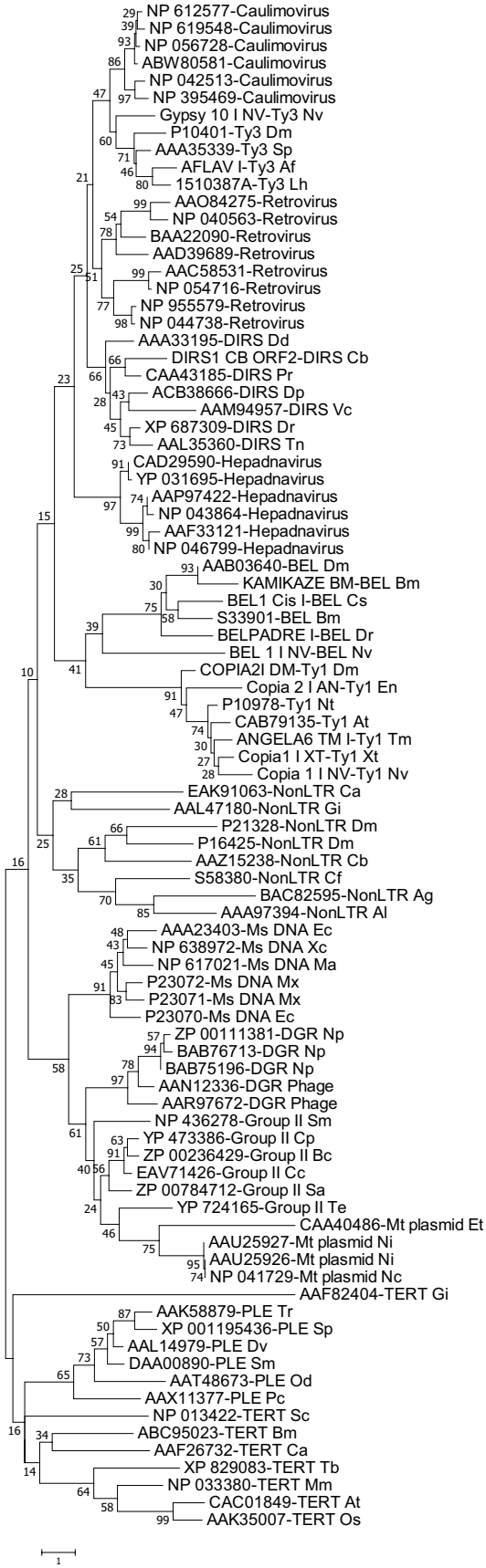


Fig. S2. (continued)

(E) **GDDA-BLAST**
RT 88

No. of Taxa : 88
Data Type : Amino acid
Analysis : Phylogeny reconstruction
Tree Inference :
->Method : Minimum Evolution
->Search Options : CNI (level = 1) with initial tree = NJ MaxTrees = 1
Phylogeny Test:
->Jacknife (1000 replicates, sampling fraction: 80%), Bootstrap (1000 replicates)

* Statistical support values presented in the order of jacknife followed by bootstrap.

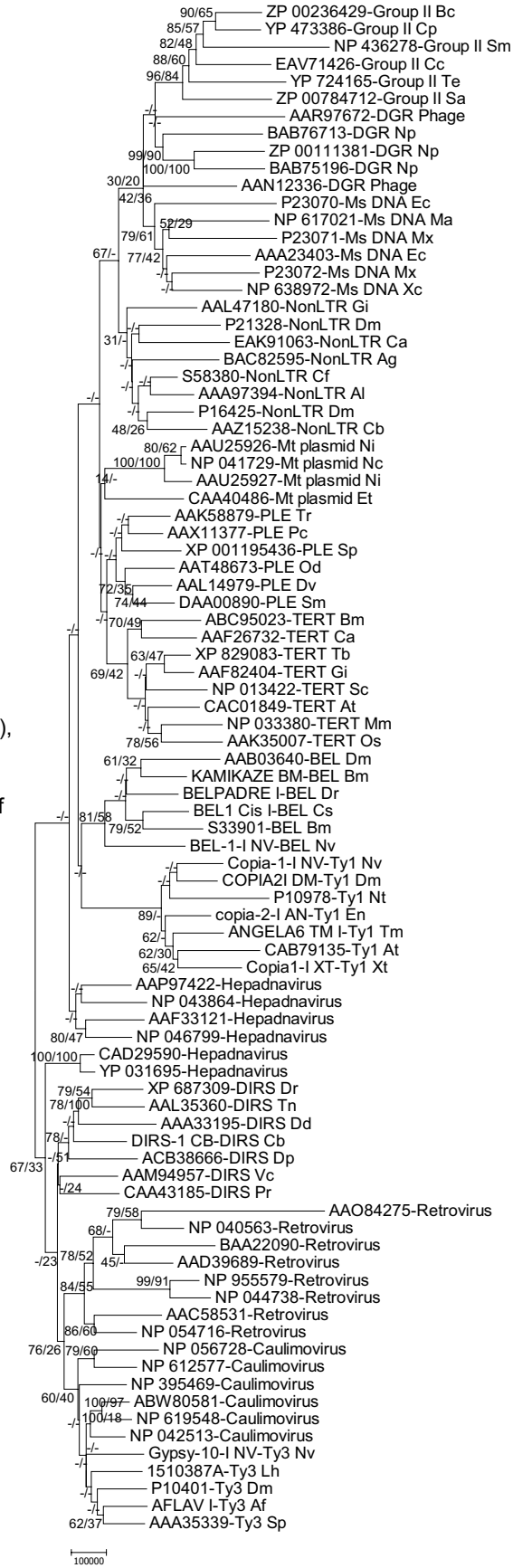


Fig. S2. (continued)

(F)

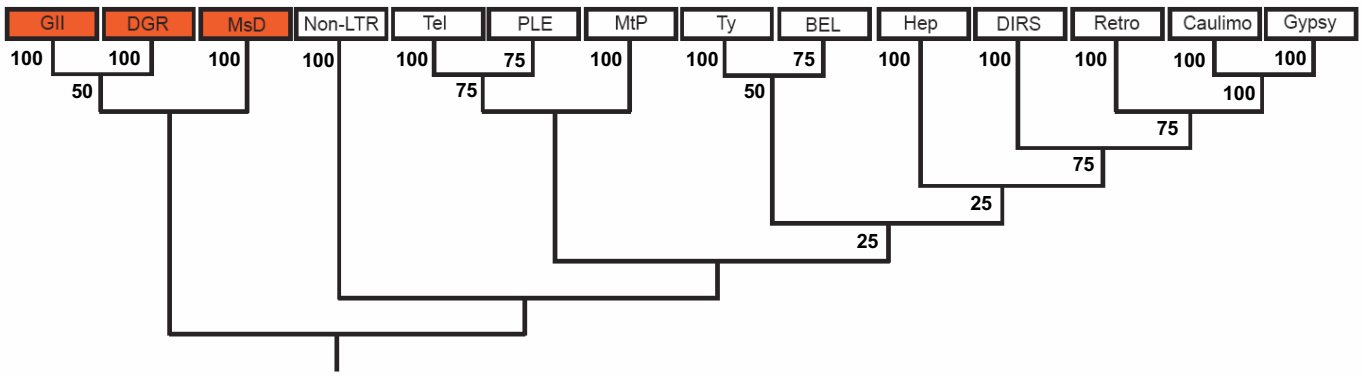


Fig. S2. (continued)

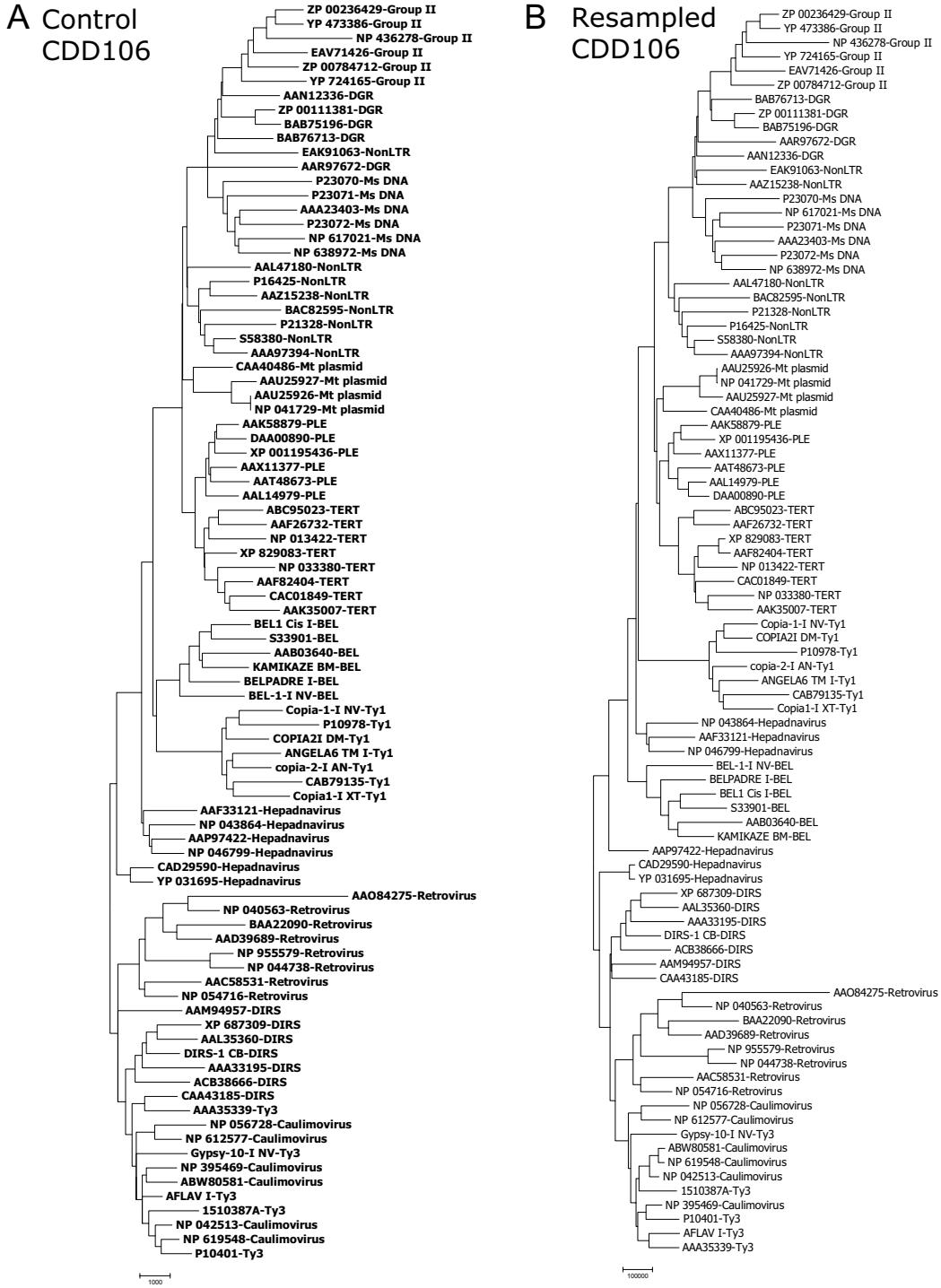
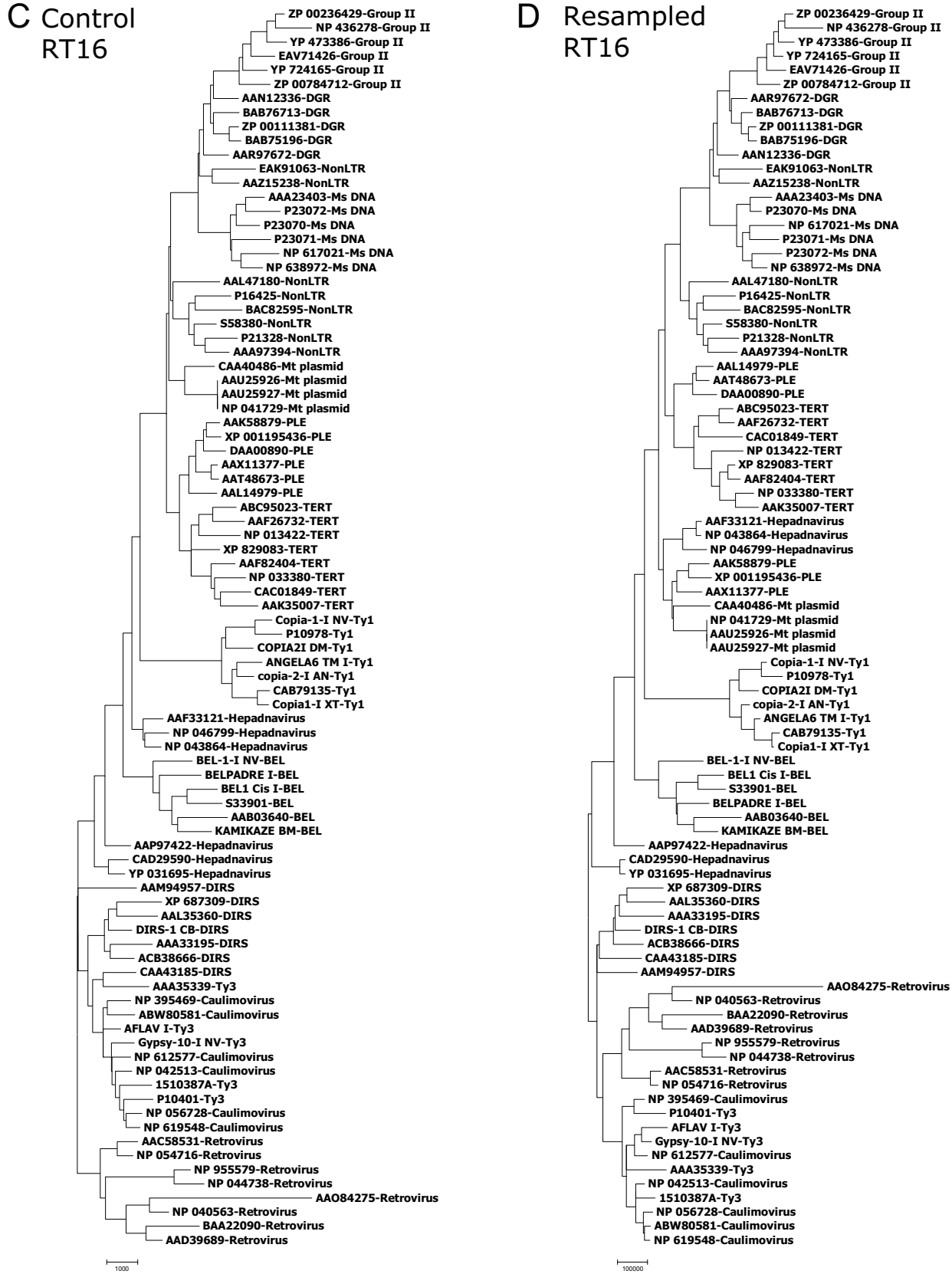


Fig. S3. Toward identifying the necessary and sufficient coordinates for GDDA-BLAST measurements. (*A* and *B*) Unrooted phylogenetic trees of the RT domain region within 88 RT sequences produced by the estimation of their evolutionary distances by GDDA-BLAST in unmodified (control) and resampled (''seeded'') conditions, respectively. Control scores are defined as mean % identity × mean % coverage, while the resampled scores include a multiplication by the Hit ratio (i.e., hit ratio = the total number of alignments to the modified query number, scaled between 0–100). Quantitative statistics of all alignments show that 107/24,280 (0.44%) profiles in the control preparation are ''active'' (having a normalized hit ratio greater than 25). (C-D) Of the 106 ''active'' profiles, the 16 RT-specific profiles are the most frequent within all 88 RT sequences. Unrooted phylogenetic trees using only these 16 profiles were generated as above. Importantly, none of the four trees presented here are as resolved as the one shown in Fig. 3. Nevertheless, these results still suggest that expanding our knowledge-base to include more of these informative profiles will increase the speed and resolution of GDDA-BLAST measurements.



Chang, GS., Hong, Y., Ko, KD., Holmes, EC, Patterson, RL., van Rossum, DB. (review, July 2008) Supplemental Figure 3

Fig. S3. (continued)

Other Supporting Information Files

Table S1
Table S2