# RepeatExplorer

## Classification of repetitive elements based on the analyzes of protein domains

**Pavel Neumann**
**May 2014**

# Classification of TEs based on protein domain search

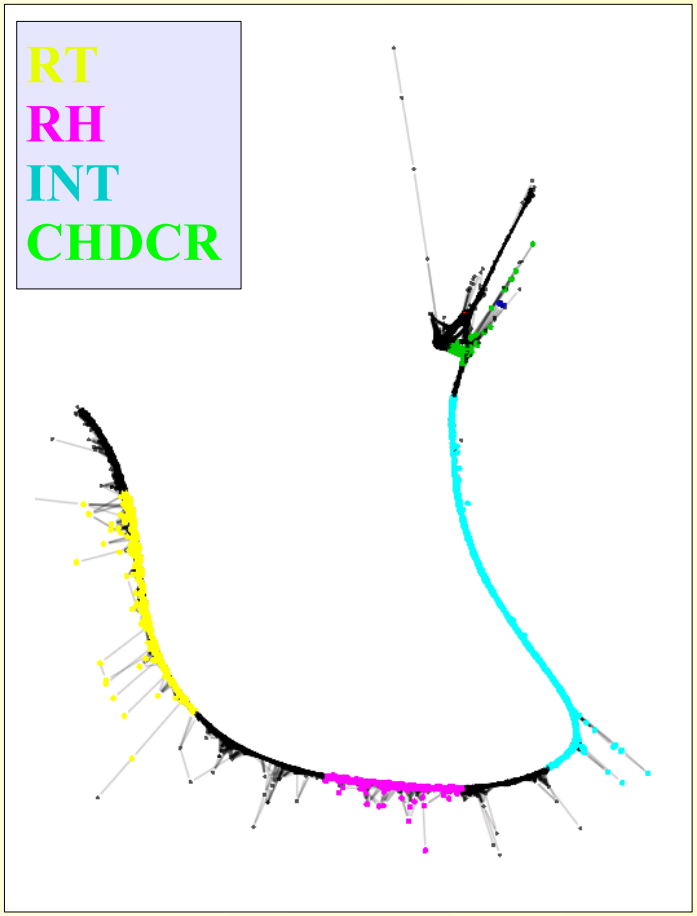
- Repetitive elements are highly diverse – their classification based on DNA sequences requires comprehensive databases which are available for a limited number of species
- Although structural features (LTR, TIR, poly-A etc.) of repetitive sequences are suitable for classification into major repeat types (e.g. LTR retrotransposons, DNA transposons, LINEs), they are usually difficult to recognize in sequence clusters produced by RepeatExplorer
- Coding sequences of transposable elements make up a large part of whole elements and encode for proteins which are relatively conserved among TEs of the same type/lineage. Therefore they are ideal for unambiguous assignment of analyzed sequences to individual types

# Preliminary classification of TEs based on protein domain search (blast using NGS reads)

HTML summary of graph based clustering

Summary of TE domain hits

| | Domain | ID | Type | Lineage | Hits | MeanScore |
|---|---|---|---|---|---|---|
| 1 | Ty3-INT | PiSat1_ID200 | Ty3/gypsy | chromovirus | 878 | 62 |
| 2 | Ty3-RT | PiSat1_ID200 | Ty3/gypsy | chromovirus | 810 | 64 |
| 3 | Ty3-CHDCR | PiSat1_ID200 | Ty3/gypsy | chromovirus | 629 | 55 |
| 4 | Ty3-RH | PiSat1_ID200 | Ty3/gypsy | chromovirus | 357 | 60 |
| 5 | Ty3-INT | MedT1_ID93 | Ty3/gypsy | chromovirus | 318 | 64 |
| 6 | Ty3-INT | MedT2_ID89 | Ty3/gypsy | chromovirus | 93 | 67 |
| 7 | Ty3-RH | LotJ2_ID65 | Ty3/gypsy | chromovirus | 67 | 63 |
| 8 | Ty3-RH | MedT2_ID89 | Ty3/gypsy | chromovirus | 66 | 71 |
| 9 | Ty3-RT | MedT2_ID89 | Ty3/gypsy | chromovirus | 54 | 70 |
| 10 | Ty3-RT | MedT1_ID93 | Ty3/gypsy | chromovirus | 43 | 64 |
| 11 | Ty3-RH | MedT1_ID93 | Ty3/gypsy | chromovirus | 11 | 51 |
| 12 | Ty3-INT | LotJ2_ID65 | Ty3/gypsy | chromovirus | 9 | 64 |
| 13 | Ty3-RT | CarP1_ID79 | Ty3/gypsy | chromovirus | 6 | 43 |
| 14 | Ty3-RT | BraR5_ID117 | Ty3/gypsy | chromovirus | 4 | 71 |
| 15 | Ty3-INT | CRR2_ID16 | Ty3/gypsy | chromovirus | 4 | 66 |
| 16 | Ty3-RT | PopT1_ID21 | Ty3/gypsy | chromovirus | 4 | 67 |
| 17 | Ty3-INT | CRA3a_ID1 | Ty3/gypsy | chromovirus | 4 | 60 |
| 18 | Ty3-CHDCR | MedT1_ID93 | Ty3/gypsy | chromovirus | 3 | 38 |
| 19 | Ty3-INT | SorB1_ID109 | Ty3/gypsy | chromovirus | 3 | 60 |



# Protein domain search: database for blast

- Protein sequences are more conserved than DNA
- **2942 domains from a total of 1045 elements**
- 398 LTR retrotransposon elements (255 Ty1/copia and 143 Ty3/gypsy)
  - 386 GAG, 394 PROT, 378 RT, 383 RNaseH, 391 INT, 31 ChDII, and 45 CHDCR domains (totally 2317domains)
- 15 pararetroviruses
  - 15 PROT, 15 RT, and 15 RNaseH domains (totally 45 domains)
- 228 nonLTR retrotransposon elements
  - 213 RT, 51 RNaseH, and 190 ENDO domains (totally 454 domains)
- 404 DNA transposons
  - totally 435domains extracted from the following transposon families: DTH (Harbinger), DTM (Mutator/MuDR), DTC (EnSpm/CACTA), DTN (Novosib), DTA (hAT/AC), DTT (Mariner), DHH (Helitron)

# Protein domain search: database for blast

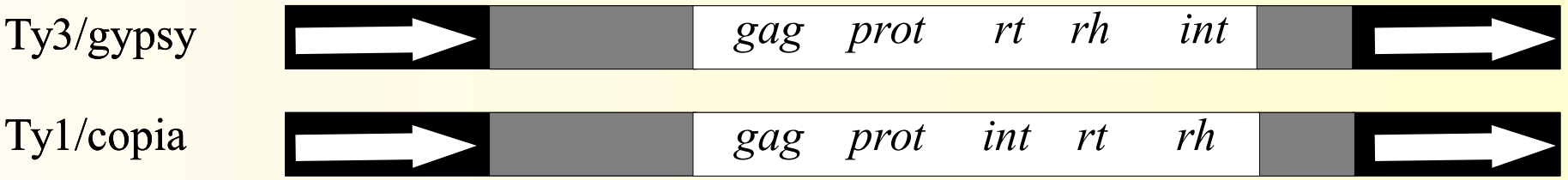
- **The database includes only plant elements, therefore, it is highly sensitive only for plant TE domains!**
- **Sources:**
  - **public databases of TEs:**
    - **RepBase (http://www.girinst.org/repbase/)**
    - **GyDB (http://gydb.org/index.php/Main_Page)**
  - **research articles:**
    - **Llorens et al., Nucleic Acids Research 2011 (mostly Ty3/gypsy)**
    - **Wicker et al., Genome Res. 2007 (Ty1/copia)**
  - **our published research:**
    - **Macas et al., Gene 2007 (Ogres)**
    - **Neumann et al., Mobile DNA 2011 (chromoviruses, CRs)**
  - **our unpublished research**
- **Protein domain sequences were extracted only from elements that seemed to have their CDS either intact or with a only a few mutations**

# Blast is suitable for a rough preliminary classification of TEs but it is not always sufficient to reliably assign novel sequences into established lineages/clades of LTR retrotransposons

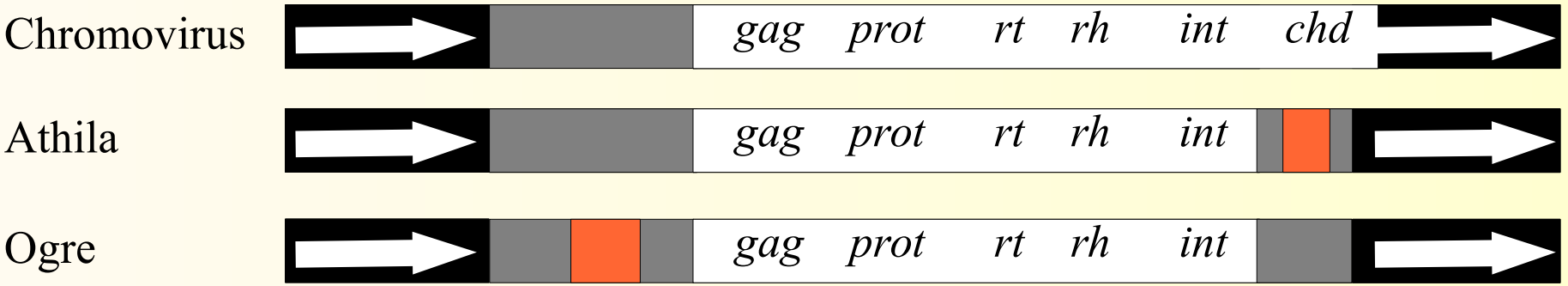
- Blast based classification is based on the best hit but the best hit is not always a strong hit

# Reliable classification of LTR retrotransposons based on the comparison of polyprotein domains

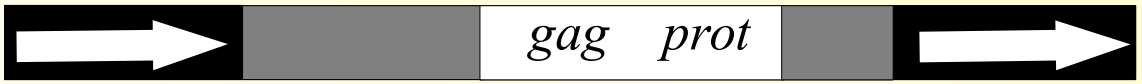
Minimal autonomous elements




Autonomous elements with extra domains or ORFs




nonautonomous elements / deletion derivatives



# RepeatExplorer: Protein domains tools

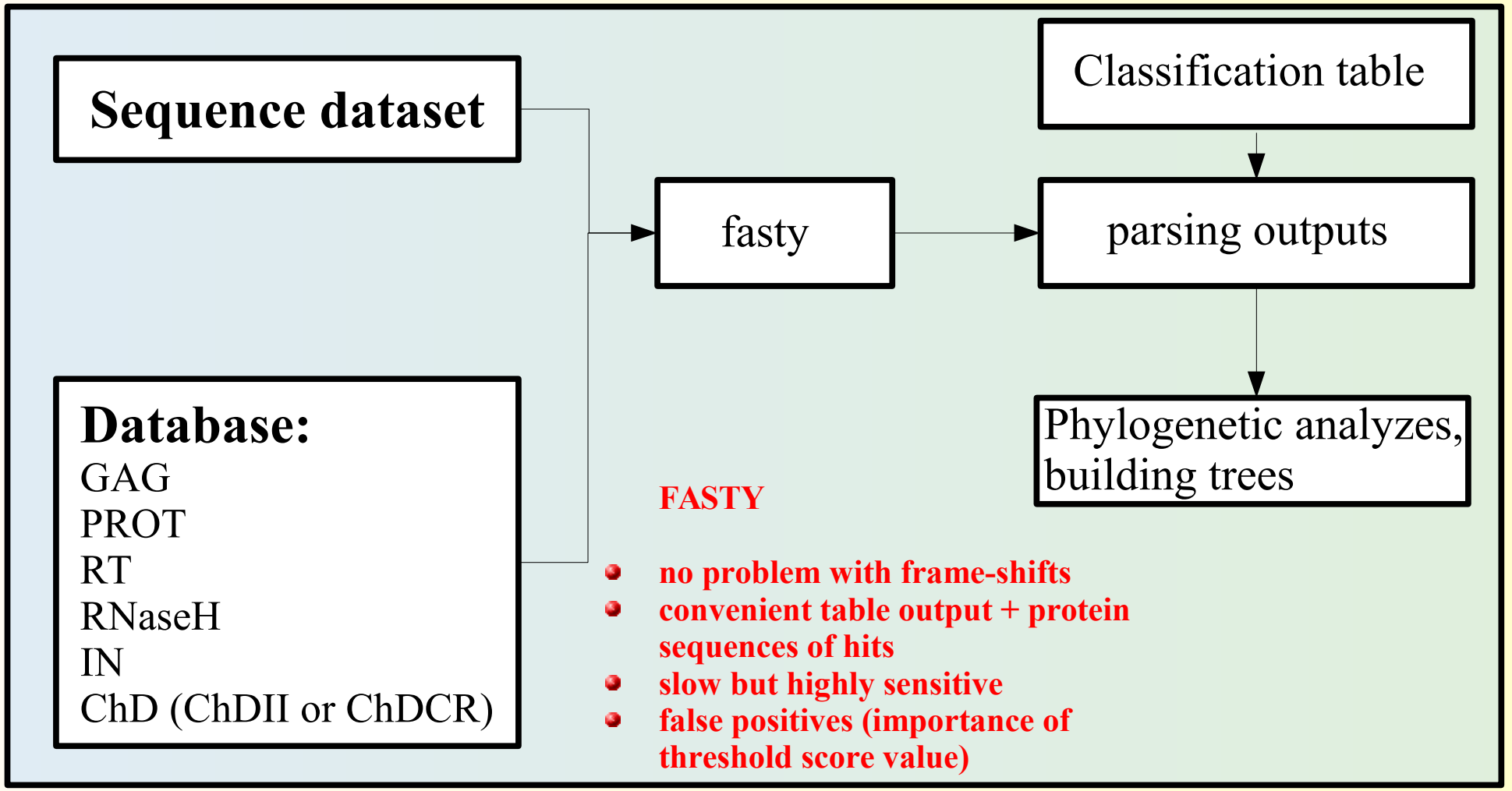
1. Protein domain search (it scans DNA sequences with protein domains database using fasty)

2. Filter output (it filters raw fasty output and extracts sequences giving hits above given threshold)

3. Create tree (it calculates and draws phylogenetic tree from a set of sequences)

# RepeatExplorer: Protein domains tools

**What it does:**



# RepeatExplorer: database and classification table

**Database:**

```
>Ty1-PROT__Copia-26_SB-I
WYMDTGATDHITSELEKLTTRDKYHGGDQVHTASGSGMEIQHIGHGVLRSPTNNLHLRNILHVPAANKDLLS
>Ty3-RT__CRM3_ID37
VRESLSPCSVPVLLVPKKDGSWRMCVDCRAINNITIRYRHPIPRLDDMLDELSGSMIFSKIDLRSGYHQIRMKLGDEWKTSFKTKFGL
YEWLVMPFGLTNAPSTFMRLMNEVLRAFIGRFVVVYFDDILIYSKCLVDHLDHLRAVFNALREARLFGNLDKCTFCTDRVSFLGY
>LINE-RH__LINE1-6_OS
DGSFQESDGKGGIGAVLRNCTGEVIFAACGHVDCCSSALETELLACRDGLALALQWTLLPIVIETDCLAMIHLFRDATGAKSELAFLI
REIDSLLVGNRDISFSKCLRSQNHISH
>DTT-CD1__Mariner-3_SBi
IDEKWFYLHQKNERYYLLPEEDVPHRTCRNKNYIPRLMFLCACARPRFRDGVCTFDGKIGCFPLVHFEPAIRSNERTGRVRGDVVMKP
ITSITGDVIRDFMINKVLPAIREKWPREDLGRPIFIQQDNA
```

**Classification table:**

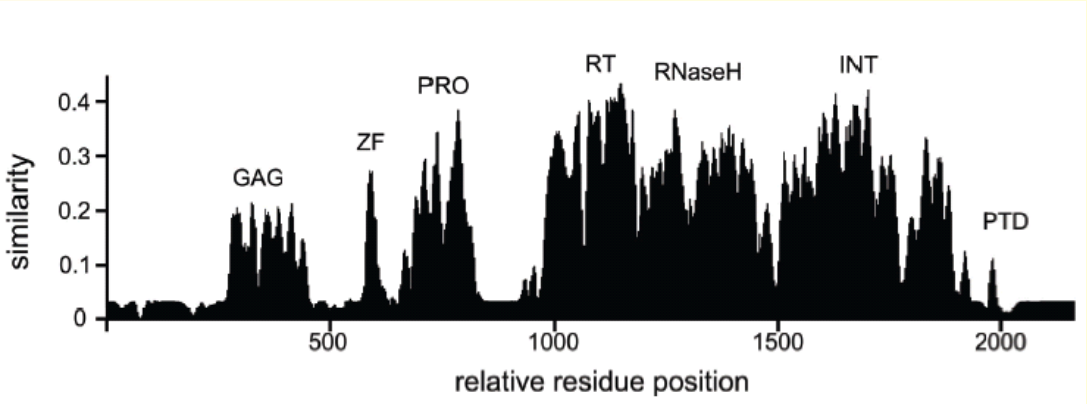
```
Copia-26_SB-I    Ty1/copia    AleI/Retrofit
CRM3_ID37    Ty3/gypsy    chromovirus
```

- **Classification table was made to classify LTR retrotransposons into lineages**
- **Types of all TEs are encoded in the name line text**

# Protein domain search: database for fasty

- 398 LTR retrotransposon elements (255 Ty1/copia and 143 Ty3/gypsy)
  - 386 GAG, 394 PROT, 378 RT, 383 RNaseH, 391 INT, 31 ChDII, and 45ChDCR domains (totally 2317domains)

- different domains have different level of divergence!



# Fasty using RT domain sequences

- RT is one of the most conserved domains of the polyprotein
- several conserved motifs are intermingled with less conserved regions and the length is highly conserved among distinct families (about 260 and 170 aa in Ty1/copia and Ty3/gypsy, respectively) - RT domain sequences are relatively easy to align
- RT domain sequences are sufficiently diverse, allowing to distinguish among distinct lineages/clades
- RT domains have already been used to infer phylogenetic relationships of retrotransposon families in plants

# Parsing and filtering output of fasty

```
fasty36 -m 3 -m 9 -b 1 -H -q -O output_file - database_file
FASTY compares a DNA sequence to a protein sequence data bank version 35.04 Aug. 25, 2009
Please cite:
 Pearson et al, Genomics (1997) 46:24-36

Query: CL46Contig6_(825-2.7-2302), 825 aa
828>>>CL46Contig6_(825-2.7-2302) - 825 aa - 825 aa
Library: all_RT_domains_for_search_extended  181897 residues in   961 sequences

 181897 residues in   961 sequences
Statistics:  Expectation_n fit: rho(ln(x))= -93.9197+/-0.0186; mu= 591.6836+/- 0.967
 mean_var=2038.9986+/-1740.653, 0's: 0 Z-trim: 325  B-trim: 0 in 0/9
 Lambda= 0.028403
Algorithm: FASTY (3.5 Aug 2009) [optimized]
Parameters: BL50 matrix (15:-5) ktup: 2
 join: 44, opt: 40, open/ext: -12/ -2 shift: -20, subs: -24 width:  16

The best scores are:                                      opt bits E(961)      %_id  %_sim   sy  alen  an0  ax0  pn0  px0  an1  ax1 pn1
px1 gapq gapl  fs
Ty3_454OCU_CL46Contig61_(1860-14.3-26665)      ( 173) [r] 1086 58.0 4.6e-10    0.931 0.971 1086  173  604   85  826 1650    1  173    1
173   0   0   1


>>>CL46Contig6_(825-2.7-2302), 825 aa vs all_RT_domains_for_search_extended library

trans. Smith-Waterman score: 1086; 93.1% identity (97.1% similar) in 173 aa overlap (604-85:1-173)
>CL46Co ..
IRPSVSPWGAPVLFVKKKDGSLRMCIDYRELNQLMVKNKYPLPRIEDLFD
QLRGAGTF\SKIDLRSGYHQLKIKECDIPKMAFRTRYGHYEFVVMPFGLS
NAPAVFMDLMNRVFHPYSDKFVIIFIDDILVYSKDKE*HQEHLRITLETL
HKEKLYAKFKKYEIWLDKVGFLGH
>Ty3_45 ..
VRPSVSPWGAPVLFVKKKDGSLRMCIDYRELNRLTVKNKYPLPRIEDLFD
QLRGAGTF-SKIDLRSGYHQLKIKECDIPKTAFRTRYGHYEFVVMPFGLS
NAPAVFMDLMNRVFHPYLDKFVIVFIDDILVYSKDKEQHQEHLRITLETL
RKEKLYAKFKKCEFWLDRVGFLGQ
```

# Parsing and filtering output of fasty: outputs

output1: table of fasty hits that passed the filter criteria

| Type | lineage | query_name | query_length | hit_name | hit_length | strand | opts | bits | e_val | perc_id | perc_sim | sw | alen | an0 | ax0 | pn0 | px0 | an1 | ax1 | pn1 | px1 | gapq | gapl | fs |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ty3/gypsy | Ogre/Tat | CL1Contig12 | 4659 | Ty3-RT__RIRE2_I | 174 | f | 878 | 243.6 | 2.20E-065 | 0.678 | 0.874 | 878 | 174 | 3035 | 3556 | 1 | 4659 | 1 | 174 | 1 | 174 | 0 | 0 | 0 |
| Ty3/gypsy | Ogre/Tat | CL1Contig13 | 4492 | Ty3-RT__RIRE2_I | 174 | f | 894 | 248 | 1.10E-066 | 0.69 | 0.885 | 894 | 174 | 3060 | 3581 | 1 | 4492 | 1 | 174 | 1 | 174 | 0 | 0 | 0 |
| Ty3/gypsy | Ogre/Tat | CL1Contig23 | 5275 | Ty3-RT__RIRE2_I | 174 | f | 878 | 243.7 | 2.50E-065 | 0.678 | 0.874 | 878 | 174 | 3301 | 3822 | 1 | 5275 | 1 | 174 | 1 | 174 | 0 | 0 | 0 |
| Ty3/gypsy | Ogre/Tat | CL1Contig26 | 3906 | Ty3-RT__RIRE2_I | 174 | f | 897 | 248.8 | 5.30E-067 | 0.695 | 0.885 | 897 | 174 | 2891 | 3412 | 1 | 3906 | 1 | 174 | 1 | 174 | 0 | 0 | 0 |
| Ty3/gypsy | Ogre/Tat | CL1Contig27 | 663 | Ty3-RT__RIRE2_I | 174 | r | 840 | 233.3 | 4.10E-063 | 0.678 | 0.879 | 840 | 174 | 584 | 61 | 663 | 1 | 1 | 174 | 1 | 174 | 0 | 0 | 2 |
| Ty3/gypsy | Ogre/Tat | CL1Contig30 | 3399 | Ty3-RT__RIRE2_I | 174 | f | 872 | 242 | 5.10E-065 | 0.672 | 0.868 | 872 | 174 | 1298 | 1819 | 1 | 3399 | 1 | 174 | 1 | 174 | 0 | 0 | 0 |
| Ty3/gypsy | Ogre/Tat | CL1Contig31 | 4756 | Ty3-RT__RIRE2_I | 174 | f | 887 | 246.1 | 4.20E-066 | 0.684 | 0.879 | 887 | 174 | 2494 | 3015 | 1 | 4756 | 1 | 174 | 1 | 174 | 0 | 0 | 0 |
| Ty3/gypsy | Ogre/Tat | CL1Contig33 | 2778 | Ty3-RT__RIRE2_I | 174 | f | 857 | 237.9 | 7.00E-064 | 0.695 | 0.885 | 857 | 174 | 2030 | 2551 | 1 | 2778 | 1 | 174 | 1 | 174 | 0 | 0 | 2 |
| Ty3/gypsy | Ogre/Tat | CL1Contig38 | 2294 | Ty3-RT__RIRE2_I | 174 | f | 870 | 241.4 | 5.00E-065 | 0.678 | 0.862 | 870 | 174 | 1592 | 2113 | 1 | 2294 | 1 | 174 | 1 | 174 | 0 | 0 | 0 |
| Ty1/copia | SIRE | CL27Contig4 | 1238 | Ty1-RT__Wicker_Maximus_cons | 258 | f | 1314 | 361.7 | 2.60E-101 | 0.888 | 0.963 | 1314 | 215 | 592 | 1236 | 1 | 1238 | 1 | 215 | 1 | 258 | 0 | 0 | 0 |
| Ty1/copia | SIRE | CL27Contig15 | 13784 | Ty1-RT__Wicker_Maximus_cons | 258 | r | 1584 | 435 | 2.50E-122 | 0.884 | 0.969 | 1584 | 258 | 5359 | 4586 | 13784 | 1 | 1 | 258 | 1 | 258 | 0 | 0 | 0 |
| Ty1/copia | SIRE | CL27Contig36 | 1114 | Ty1-RT__Wicker_Maximus_cons | 258 | r | 1549 | 425.4 | 1.50E-120 | 0.872 | 0.957 | 1549 | 258 | 967 | 194 | 1114 | 1 | 1 | 258 | 1 | 258 | 0 | 0 | 0 |
| Ty1/copia | Angela | CL55Contig6 | 3043 | Ty1-RT__Wicker_BARE1_B_cons | 258 | f | 1586 | 435.5 | 3.90E-123 | 0.933 | 0.992 | 1586 | 255 | 2 | 766 | 1 | 3043 | 4 | 258 | 1 | 258 | 0 | 0 | 0 |
| Ty1/copia | TAR | CL89Contig4 | 5745 | Ty1-RT__Wicker_Osr1_Os_cons | 257 | r | 1582 | 434.4 | 1.50E-122 | 0.891 | 0.973 | 1582 | 257 | 4461 | 3691 | 5745 | 1 | 1 | 257 | 1 | 257 | 0 | 0 | 0 |

```
>CL27Contig4_592-1236
WIQAMQEELQQFELNNVWELVKRPDPRKHNIIGTKWIYRNKQDEHGQVVR
NKARLVAQGYTQVEGIDFDETFAPVARLEAIRILLAYANHHNIFLYQMDV
KSAFLNGKIEEEVYVAQPPGFEDPKRPDMVYKLNKALYGLKQAPRAWYDT
LKDFLKSKGFKPGSLDPTLFTKTYDGELFVCQIYVDDIIFGCTVKRYSYE
FGYMMQVQYQMSMMG
>CL27Contig15_5359-4586
WIQAMQEELQQFELNNVWELVKRPDPRKHNIIGTKWIYRNKQDEHGQVVR
NKARLVAQGYTQVEGIDFDETFAPVARLEAIRILLAYANHHNILLYQMDV
KSAFLNGRIEEEVYVAQPPGFEDPKRPDMVYKLNKALYGLKQAPRAWYDT
LKDFLKSKGFKPGSLDPTLFTKTYDDELFVCQIYVDDIIFGCTDKRYSDE
FGHMMQEQYQMSMMGELKFFLGLQIRQQSNGIFISQEKYLKDCLKKFGMQ
DCKGYSTP
>CL27Contig36_967-194
WIQAMHEELQQFELNNVWELVKRPDPRKHNIIGTKWIYRNKQDENGQVVI
NKARLVAQGYTQVEGIDFDETFALVARLEAIRILLAYANHHNILLYQMDV
KSAFLNGKIEEEVYVAQPPGFEDPKRPDMVYKLNKALYGLKQAPRAWYDT
LKDFLKSKGFKPGSLDPTLFTKTYDDELFVCQIYVDDIIFGCTDKRYSDE
FGHMMQEQYQMSMMGELKFFLGLQIRQQSNDIFISQEKYLKDCLKKFGMQ
DCKGYTTP
```

output2: Ty1 domains

```
>CL1Contig12_3035-3556
VRGVIHPTWLANPVVVRKANGKWRLCIDFTDVNKACPKDPFPLPRIDQIV
DSTAGCELLSFLDAYSGYHQIFMAKEDEEKTAFITPCGTYCFIRMPFGLK
NAGSTFARVVYTAFEPQIHRNVEAYMDDIVVKSKSKETLIQDLEETFQNL
RRIQLKLNPEKCVFGVPSGKLLGF
>CL1Contig13_3060-3581
VRGVIHPTWLANPVVVRKANGKWRLCIDFTDVNKACPKDPFPLPRIDQIV
DSTAGCELLSFLDAYSGYHQIFMAEEDEEKTAFITPCGTYCFIRMPFGLK
NAGSTFARVVHTAFEPQIHRNVEAYMDDIVVKSKDKDTLIQDLDETFANL
RKIQLKLNPEKCVFGVPSGKLLGF
>CL1Contig23_3301-3822
VRGVLHPTWLANPVVVRKANGKWRLCIDFTDVNKACPKDPFPLPRIDQIV
DSTAGCDLLSFLDAYSGYHQIFMAEEDEEKTAFITPCGTYCFIRMPFGLK
NAGSTFARVVHTAFEPQIHRNVEAYMDDIVVKSKDRATLIQDLDETFANL
RKINLKLNPEKCVFGVPSGKLLGF
>CL1Contig26_2891-3412
VRGVLHPTWLANPVVVRKANGKWRLCIDFTDVNKACPKDPFPLPRIDQIV
DSTAGCELLSFLDAYSGYHQIFMAEEDEEKTAFITPCGTYCFIRMPFGLK
NAGSTFARVVHTAFEPQIHRNVEAYMDDIVVKSKDKDTLIQDLDETFANL
RKIQLKLNPEKCVFGVPSGKLLGF
```

output3: Ty3 domains

# Parsing and filtering output of fasty: outputs

- Number of identified full-length domains depends on the length of NGS reads (the longer the reads the higher number of full-length domains)

- Fasty is run with default settings - the stringency is very relaxed which leads to many false positives

- False positives can be easily removed at the filtering step but keep in mind that different domains require different values of key parameters
  - minimum similarity [%]: **60**
  - minimum identity [%]: **40**
  - proportion of the hit length from the length of the database sequence [0-1]: **0.8**

# Classification of LTR retrotransposons: table or tree?

output1: table of fasty hits that passed the filter criteria

| Type | lineage | query_name | query_length | hit_name | hit_length | strand | opts | bits | e_val | perc_id | perc_sim | sw | alen | an0 | ax0 | pn0 | px0 | an1 | ax1 | pn1 | px1 | gapq | gapl | fs |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ty3/gypsy | Ogre/Tat | CL1Contig12 | 4659 | Ty3-RT__RIRE2_I | 174 | f | 878 | 243.6 | 2.20E-065 | 0.678 | 0.874 | 878 | 174 | 3035 | 3556 | 1 | 4659 | 1 | 174 | 1 | 174 | 0 | 0 | 0 |
| Ty3/gypsy | Ogre/Tat | CL1Contig13 | 4492 | Ty3-RT__RIRE2_I | 174 | f | 894 | 248 | 1.10E-066 | 0.69 | 0.885 | 894 | 174 | 3060 | 3581 | 1 | 4492 | 1 | 174 | 1 | 174 | 0 | 0 | 0 |
| Ty3/gypsy | Ogre/Tat | CL1Contig23 | 5275 | Ty3-RT__RIRE2_I | 174 | f | 878 | 243.7 | 2.50E-065 | 0.678 | 0.874 | 878 | 174 | 3301 | 3822 | 1 | 5275 | 1 | 174 | 1 | 174 | 0 | 0 | 0 |
| Ty3/gypsy | Ogre/Tat | CL1Contig26 | 3906 | Ty3-RT__RIRE2_I | 174 | f | 897 | 248.8 | 5.30E-067 | 0.695 | 0.885 | 897 | 174 | 2891 | 3412 | 1 | 3906 | 1 | 174 | 1 | 174 | 0 | 0 | 0 |
| Ty3/gypsy | Ogre/Tat | CL1Contig27 | 663 | Ty3-RT__RIRE2_I | 174 | r | 840 | 233.3 | 4.10E-063 | 0.678 | 0.879 | 840 | 174 | 584 | 61 | 663 | 1 | 1 | 174 | 1 | 174 | 0 | 0 | 2 |
| Ty3/gypsy | Ogre/Tat | CL1Contig30 | 3399 | Ty3-RT__RIRE2_I | 174 | f | 872 | 242 | 5.10E-065 | 0.672 | 0.868 | 872 | 174 | 1298 | 1819 | 1 | 3399 | 1 | 174 | 1 | 174 | 0 | 0 | 0 |
| Ty3/gypsy | Ogre/Tat | CL1Contig31 | 4756 | Ty3-RT__RIRE2_I | 174 | f | 887 | 246.1 | 4.20E-066 | 0.684 | 0.879 | 887 | 174 | 2494 | 3015 | 1 | 4756 | 1 | 174 | 1 | 174 | 0 | 0 | 0 |
| Ty3/gypsy | Ogre/Tat | CL1Contig33 | 2778 | Ty3-RT__RIRE2_I | 174 | f | 857 | 237.9 | 7.00E-064 | 0.695 | 0.885 | 857 | 174 | 2030 | 2551 | 1 | 2778 | 1 | 174 | 1 | 174 | 0 | 0 | 2 |
| Ty3/gypsy | Ogre/Tat | CL1Contig38 | 2294 | Ty3-RT__RIRE2_I | 174 | f | 870 | 241.4 | 5.00E-065 | 0.678 | 0.862 | 870 | 174 | 1592 | 2113 | 1 | 2294 | 1 | 174 | 1 | 174 | 0 | 0 | 0 |
| Ty1/copia | SIRE | CL27Contig4 | 1238 | Ty1-RT__Wicker_Maximus_cons | 258 | f | 1314 | 361.7 | 2.60E-101 | 0.888 | 0.963 | 1314 | 215 | 592 | 1236 | 1 | 1238 | 1 | 215 | 1 | 258 | 0 | 0 | 0 |
| Ty1/copia | SIRE | CL27Contig15 | 13784 | Ty1-RT__Wicker_Maximus_cons | 258 | r | 1584 | 435 | 2.50E-122 | 0.884 | 0.969 | 1584 | 258 | 5359 | 4586 | 13784 | 1 | 1 | 258 | 1 | 258 | 0 | 0 | 0 |
| Ty1/copia | SIRE | CL27Contig36 | 1114 | Ty1-RT__Wicker_Maximus_cons | 258 | r | 1549 | 425.4 | 1.50E-120 | 0.872 | 0.957 | 1549 | 258 | 967 | 194 | 1114 | 1 | 1 | 258 | 1 | 258 | 0 | 0 | 0 |
| Ty1/copia | Angela | CL55Contig6 | 3043 | Ty1-RT__Wicker_BARE1_B_cons | 258 | f | 1586 | 435.5 | 3.90E-123 | 0.933 | 0.992 | 1586 | 255 | 2 | 766 | 1 | 3043 | 4 | 258 | 1 | 258 | 0 | 0 | 0 |
| Ty1/copia | TAR | CL89Contig4 | 5745 | Ty1-RT__Wicker_Osr1_Os_cons | 257 | r | 1582 | 434.4 | 1.50E-122 | 0.891 | 0.973 | 1582 | 257 | 4461 | 3691 | 5745 | 1 | 1 | 257 | 1 | 257 | 0 | 0 | 0 |

- Classification included in the table is reliable only in case of high similarity to the database sequences
- Sequences sharing low similarity with those from database should be classified using phylogenetic analysis

# Classification of LTR retrotransposons using phylogenetic analysis of their domains

- it helps to assign retrotransposon sequences into phylogenetic lineages which were established previously
- trees are calculated using NJ algorithm because its fast and provides reasonably accurate results
- inferring detailed phylogenetic relationships among identified elements should be done using more sophisticated algorithms and models

# Need for unified classification of TEs including LTR retrotransposons

- Nomenclature, and classification of TEs into major types is well established but there are no rules for more detailed classification into subclasses/lineages/families ….
- The classification should be hierarchical, following principals similar to those used for taxonomy of species.
    - Results of different studies on LTR retrotransposons are difficult to compare because relationships among described sequences is often not clear (there are a number of papers describing similar TEs under different names)
- The deeper the level of classification the bigger the problem of category definition (continuum of sequence similarity)
    - e.g. 80-80-80 rule for family definition
- Some attempts have been made but it is still far from perfect
    - Wicker et al. 2007: Class – subclass – order -superfamily -family – subfamily -insertion

GUIDELINES

# A unified classification system for eukaryotic transposable elements

*Thomas Wicker, François Sabot, Aurélie Hua-Van, Jeffrey L. Bennetzen, Pierre Capy, Boulos Chalhoub, Andrew Flavell, Philippe Leroy, Michele Morgante, Olivier Panaud, Etienne Paux, Phillip SanMiguel and Alan H. Schulman*

Abstract | Our knowledge of the structure and composition of genomes is rapidly progressing in pace with their sequencing. The emerging data show that a significant portion of eukaryotic genomes is composed of transposable elements (TEs). Given the abundance and diversity of TEs and the speed at which large quantities of sequence data are emerging, identification and annotation of TEs presents a significant challenge. Here we propose the first unified hierarchical classification system, designed on the basis of the transposition mechanism, sequence similarities and structural relationships, that can be easily applied by non-experts. The system and nomenclature is kept up to date at the WikiPoson web site.

Our knowledge of the structure and composition of genomes has progressed in pace with their sequencing. As was expected early on[1], TEs have been found in virtually all eukaryotic species investigated so far[2–4]. The only known exceptions are *Plasmodium falciparum* and probably several closely related species. The TEs affect the genome by their ability to move and replicate, thereby generating plasticity. Although the main TE groups are ancient and are present in all kingdoms, TEs display extreme diversity: there are thousands or even tens of thousands of different TE families in plants[5–8], constituting 80% or more of the total genomic DNA. Although they are less abundant than those in plants, TEs in fungi and metazoans also represent a substantial part of their genomes (3–20% in fungi and 3–45% in metazoans), and include members of most superfamilies[9,10].

More and more large and complex eukaryotic genomes are being selected for sequencing, and they too are expected to be rich in TEs. Given the abundance and diversity of TEs, these genomes will present an enormous TE identification and annotation problem. Hence, there is an increasing demand for tools to handle these rapidly emerging sequences effectively. This requires straightforward and efficient nomenclature keys and classification strategies that can help non-specialists to easily annotate TEs.

The sequencing of the genomes of multiple related species will facilitate comparative studies and provide insights into genome evolution. However, comparisons will be compromised unless TE annotation is consistent. Until now, TE analysis and classification have generally been carried out on a species-by-species basis, and comparative studies of TEs between taxa have been rare. Thus, no unified system has been applied across species and, therefore, across kingdoms. Accordingly, TE databases are generally restricted to individual or closely related species and tend to lack systematic structure.

Here we propose a common TE classification system that can be easily handled by non-specialists during annotation. It provides a consensus between the various conflicting classification and naming systems that are currently in use. A key component of this system is a naming convention: a three-letter code with each letter respectively denoting class, order and superfamily; the family (or subfamily) name; the sequence (database accession) on which the element was found; and the 'running number', which defines the individual insertion in the accession. The unified system is also intended to facilitate comparative and evolutionary studies on TEs from different species.

## Outline of the classification system

In 1989, Finnegan proposed the first TE classification system, which distinguished two classes by their transposition intermediate: RNA (class I or retrotransposons) or DNA (class II or DNA transposons). The transposition mechanism of class I is commonly called 'copy-and-paste', and that of class II, 'cut-and-paste'[11]. The discovery of bacterial[12] and eukaryotic[13,14] TEs that copy and paste but without RNA intermediates, and of highly reduced non-autonomous TEs called miniature inverted repeat transposable elements (MITEs), has challenged the two-class system. These findings led to schemes that either invoke a third class or jettison the two-class system in favour of enzymological categories[15].

Our hierarchical TE classification system (FIG. 1) reconciles both approaches, maintaining two classes while applying mechanistic and enzymatic criteria. It includes (in hierarchical order) the levels of class, subclass, order, superfamily, family and subfamily. These terms were chosen to mirror those used in organismal phylogenetics and to recognize the fact that some terms (such as superfamily) are already in widespread use for TE groups. The highest level (class) divides TEs by the presence or absence of an RNA transposition intermediate, as before[11]. Subclass, previously used to separate long terminal repeat (LTR) from non-LTR (long and short interspersed nuclear element (LINE and SINE)) class I TEs, here is used to distinguish elements that copy themselves for insertion from those that leave the donor site to reintegrate elsewhere. It concomitantly reflects the number of DNA strands that are cut at the TE donor site. A new taxon — order — marks major differences in the insertion mechanism and, consequently, overall organization and enzymology, thereby replacing subclass for type I TEs.







# A unified classification system for eukaryotic transposable elements should reflect their phylogeny

*Ole Seberg and Gitte Petersen*

To assist genome annotators in naming transposable elements (TEs), Wicker *et al.* propose in their Guidelines article (A unified classification system for eukaryotic transposable elements. *Nature Rev. Genet.* 8, 973–982 (2007))[1] a classification and nomenclatural system aimed at simplifying annotation to assist future detailed structural, functional and evolutionary analyses[1]. However, even though the proposed classification system mimics organismal phylogenetic classification, it fails to do so consistently.

Classifications are a necessary prerequisite for human communication, and most biological classifications are hierarchical. The relationships between entities in any hierarchical system are of two kinds only: either one entity is included in another or they are mutually exclusive. The only rule that applies to the ranks in biological nomenclature is their sequence — a class includes one or more orders, an order one or more families and so on. Most biologists agree that classifications should reflect phylogeny (see REF. 2 for one exception), although how phylogeny is reconstructed is controversial. This aside, the majority of organismal phylogenies are inherently hierarchical, and thus either a clade is included in another clade or two clades are mutually exclusive. Hence, as long as the groups remain monophyletic[3] (for an alternative point of view see REFS 4,5), a hierarchical classification is ideally suited to reflect phylogeny[4].

Some groups with formal rank in the TE classification proposed by Wicker *et al.*[1] seem to be well justified (for example, classes). However, the limits of groups become fuzzier as one descends through the hierarchy and, to an increasing degree, loses connection with phylogeny. Some groups are surely not monophyletic; for example, autonomous and non-autonomous partners of the same family are classified as different subfamilies. In addition, although recognizing non-monophyly of miniature inverted repeat elements (MITEs), Wicker *et al.*[1] repeat the classification error by recognition of paraphyletic subfamilies that include only autonomous members, even though they claim that "subfamilies are defined on the basis of phylogenetic data"[1]. Furthermore, the 80–80–80 similarity rule that is used to create families and subfamilies leads to an unpredictable mix of monophyletic, paraphyletic and polyphyletic groups. Naming paraphyletic groups may be misleading[3], but naming polyphyletic groups has no connection to reality.

Additionally, it must be remembered that, despite their mobility, TEs over time 'travel' within their host genome. Hence, although a TE phylogeny corresponds to a gene tree and need not reflect species phylogeny, the evolutionary histories of the TEs and of their hosts are tightly linked, and will need to be integrated to achieve a full understanding of both[7].

It is particularly discouraging that Wicker *et al.*[1], despite being aware of the fact that homology is explained by common descent, misuse[8,9] the concept by reference to "a continuum of sequence homology"[1]. Homology (as with pregnancy, for example) is not measured in degrees: either two sequences are homologous or not. However, in contrast to pregnancy, homology is only a hypothesis[8,10–12].

We absolutely accept the need for a classification system of TEs, but equally urge that such a system must be based on the consistent use of principles. Anything else may retard, rather than improve, our understanding of TE evolution.

*Ole Seberg and Gitte Petersen are at the Laboratory of Molecular Systematics, Natural History Museum of Denmark, Sølvgade 83, opg. S, DK-1307 Copenhagen K, Denmark.*

Published online 24 February 2009

1. Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. *Nature Rev. Genet.* 8, 973–982 (2007).
2. Felsenstein, J. *Inferring Phylogenies* (Sinauer Associates, Sunderland, Massachusetts, 2004).
3. Nelson, G., Murphy, D. J. & Ladiges, P. Y. Brummitt on paraphyly: a response. *Taxon* 52, 295–298 (2003).
4. Brummitt, R. K. Further dogged defense of paraphyletic taxa. *Taxon* 52, 803–804 (2003).
5. Nordal, I. & Stedje, B. Paraphyletic taxa should be accepted. *Taxon* 54, 5–8 (2005).
6. Sober, E. *Philosophy of Biology* (Oxford Univ. Press, Oxford, 1993).
7. Petersen, G. & Seberg, O. Stowaway MITEs in Hordeum (Poaceae): evolutionary history, ancestral elements, and classification. *Cladistics* (in the press).
8. Patterson, C. Homology in classical and molecular biology. *Mol. Biol. Evol.* 5, 603–625 (1988).
9. Swofford, D. L., Olsen, G. J., Waddell, P. J. & Hillis, D. M. in *Molecular Systematics* (eds Hillis, D. M., Moritz, C. & Mable, B. K.) 407–514 (Sinauer Associates Inc., Sunderland, Massachusetts, 1996).
10. de Pinna, M. C. C. Concepts and tests of homology in the cladistic paradigm. *Cladistics* 7, 367–394 (1991).
11. Brower, A. V. Z. & Schawaroch, V. Three steps of homology assessment. *Cladistics* 12, 265–272 (1996).
12. Williams, D. M. in *Milestones in Systematics* (eds Williams, D. M. & Forey, P. L.) 191–224 (CRC, Boca Raton, Florida 2004).



# Reply: A unified classification system for eukaryotic transposable elements should reflect their phylogeny

*Thomas Wicker, François Sabot, Aurélie Hua-Van, Jeffrey L. Bennetzen, Pierre Capy, Boulos Chalhoub, Andrew Flavell, Philippe Leroy, Michele Morgante, Olivier Panaud, Etienne Paux, Phillip SanMiguel and Alan H. Schulman*

# A unified classification system for eukaryotic transposable elements

*Thomas Wicker, François Sabot, Aurélie Hua-Van, Jeffrey L. Bennetzen, Pierre Capy, Boulos Chalhoub, Andrew Flavell, Philippe Leroy, Michele Morgante, Olivier Panaud, Etienne Paux, Phillip SanMiguel and Alan H. Schulman*

Table 2 | **Examples of the application of the proposed classification system**

| | *Barbara* | *Sukkula* | *Thalos* | *Xithos* |
|---|---|---|---|---|
| **Class** | Retrotransposon | Retrotransposon | DNA transposons | Unknown |
| **Subclass** | N/A | N/A | 1 | Unknown |
| **Order** | LTR retrotransposon | LTR retrotransposon | TIR | Unknown |
| **Superfamily** | *Copia* | Unknown | *Mariner* | Unknown |
| **Family** | *Barbara* | Unknown | *Stowaway* | *Xithos* |
| **Subfamily** | Not defined | *Sukkula* | *Thalos* | Not defined |
| **Insertion** | RLC_*Barbara*_AF1234-2 | RLx_*Sukkula*_AF1234-1 | DTT_*Thalos*_AF1234-8 | xxx_Xithos_AF1234-1 |
| **Structural description** | Autonomous retrotransposon | LARD | MITE | - |

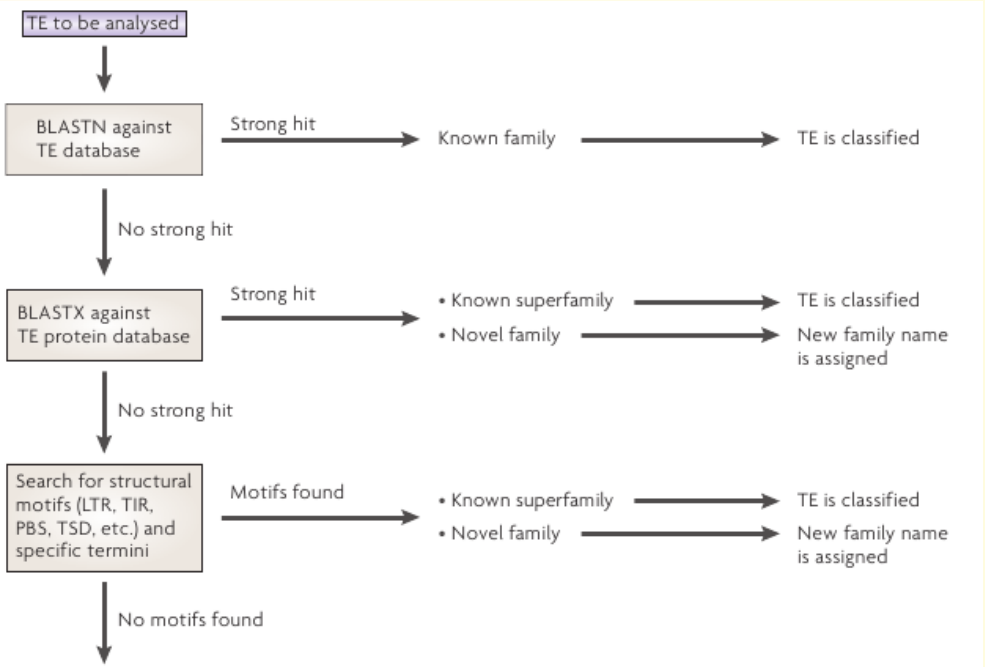


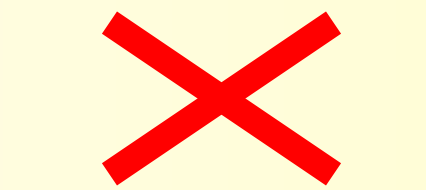

Figure 3 | **Step by step transposable element (TE) classification.** LTR, long terminal repeat; PBS, primer binding site; TIR, terminal inverted repeat; TSD, target site duplication.

# A unified classification system for eukaryotic transposable elements should reflect their phylogeny

*Ole Seberg and Gitte Petersen*

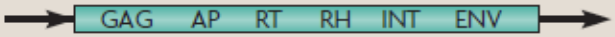
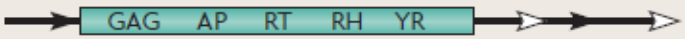
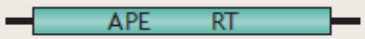
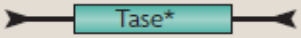
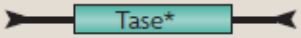
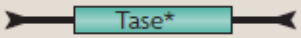
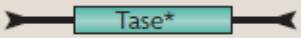
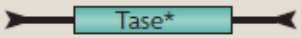
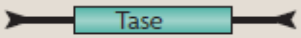
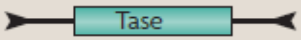
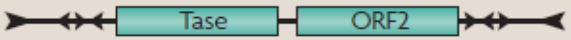
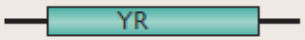
We absolutely accept the need for a classification system of TEs, but equally urge that such a system must be based on the consistent use of principles. Anything else may retard, rather than improve, our understanding of TE evolution.

**WikiPoson web site**



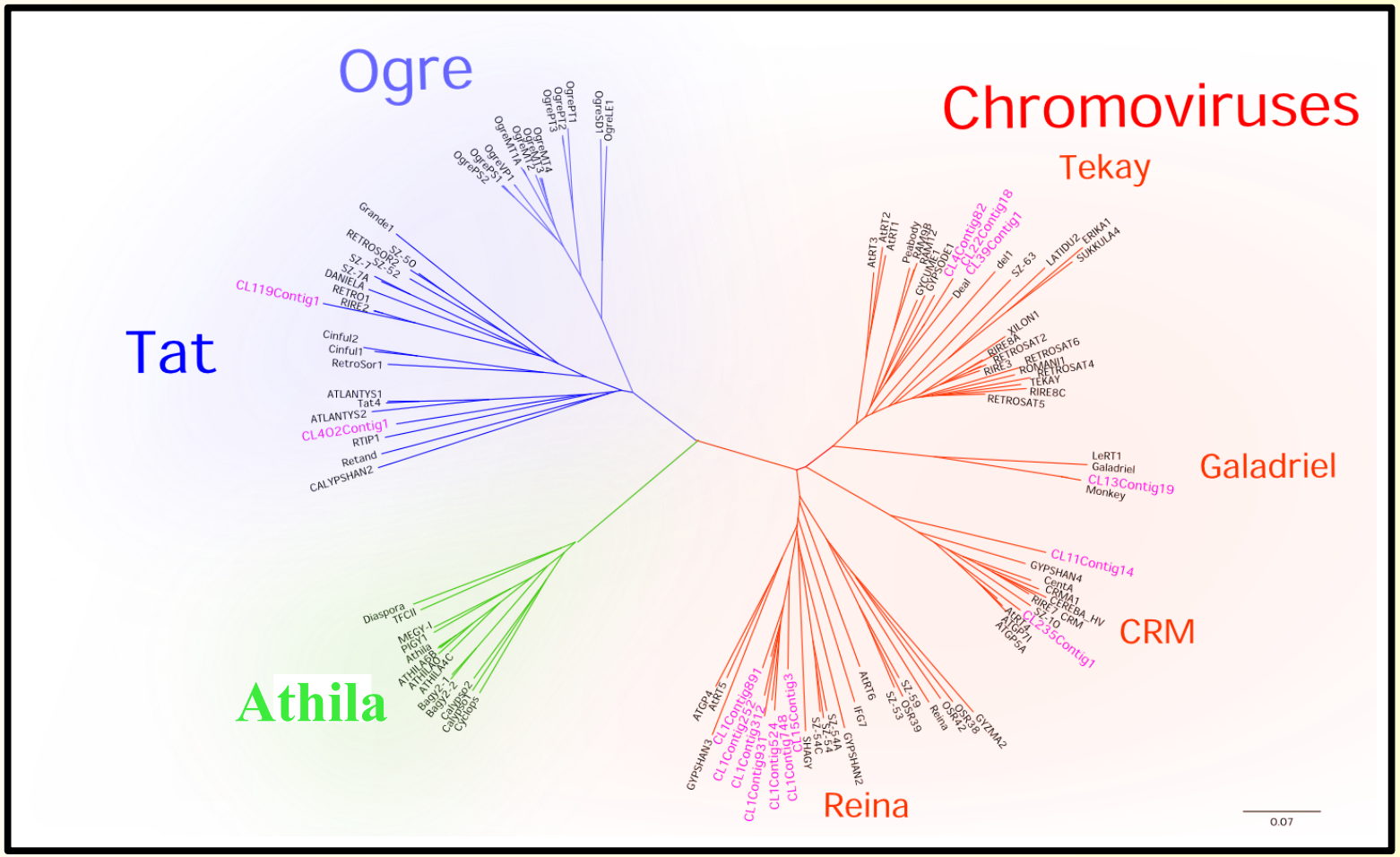

*navigation*

- *Starting page*
- *Classification Basics*
- *The Classification itself*
- *Elements currently described here*
- *Nomenclature Basics*
- *Species and TEs*
- *TE proteins*

| Classification | | Structure | TSD | Code | Occurrence |
|---|---|---|---|---|---|
| **Order** | **Superfamily** | | | | |
| *Class I (retrotransposons)* | | | | | |
| LTR | *Copia* |   | 4–6 | RLC | P, M, F, O |
| | *Gypsy* |   | 4–6 | RLG | P, M, F, O |
| | *Bel–Pao* |   | 4–6 | RLB | M |
| | *Retrovirus* |   | 4–6 | RLR | M |
| | *ERV* |   | 4–6 | RLE | M |
| DIRS | *DIRS* |   | 0 | RYD | P, M, F, O |
| | *Ngaro* |   | 0 | RYN | M, F |
| | *VIPER* |   | 0 | RYV | O |
| PLE | *Penelope* |   | Variable | RPP | P, M, F, O |
| LINE | *R2* |   | Variable | RIR | M |
| | *RTE* |   | Variable | RIT | M |
| | *Jockey* |   | Variable | RIJ | M |
| | *L1* |   | Variable | RIL | P, M, F, O |
| | *I* |   | Variable | RII | P, M, F |
| SINE | *tRNA* |  | Variable | RST | P, M, F |
| | *7SL* |  | Variable | RSL | P, M, F |
| | *5S* |  | Variable | RSS | M, O |
| *Class II (DNA transposons) - Subclass 1* | | | | | |
| TIR | *Tc1–Mariner* |   | TA | DTT | P, M, F, O |
| | *hAT* |   | 8 | DTA | P, M, F, O |
| | *Mutator* |   | 9–11 | DTM | P, M, F, O |
| | *Merlin* |   | 8–9 | DTE | M, O |
| | *Transib* |   | 5 | DTR | M, F |
| | *P* |   | 8 | DTP | P, M |
| | *PiggyBac* |   | TTAA | DTB | M, O |
| | *PIF–Harbinger* |   | 3 | DTH | P, M, F, O |
| | *CACTA* |   | 2–3 | DTC | P, M, F |
| Crypton | *Crypton* |   | 0 | DYC | F |
| *Class II (DNA transposons) - Subclass 2* | | | | | |
| Helitron | *Helitron* |   | 0 | DHH | P, M, F |
| Maverick | *Maverick* |   | 6 | DMM | M, F, O |

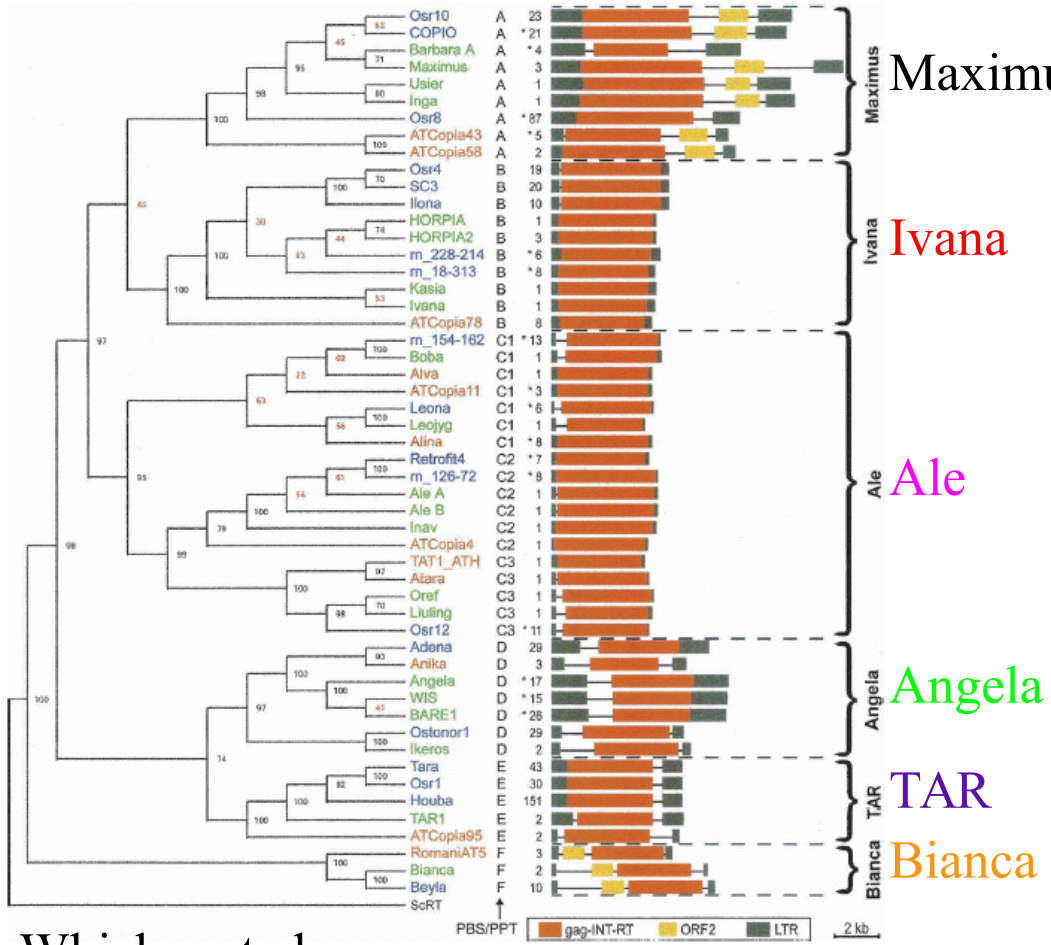
# Ty3/gypsy

## modified classification according to Llorens et al. (2011)

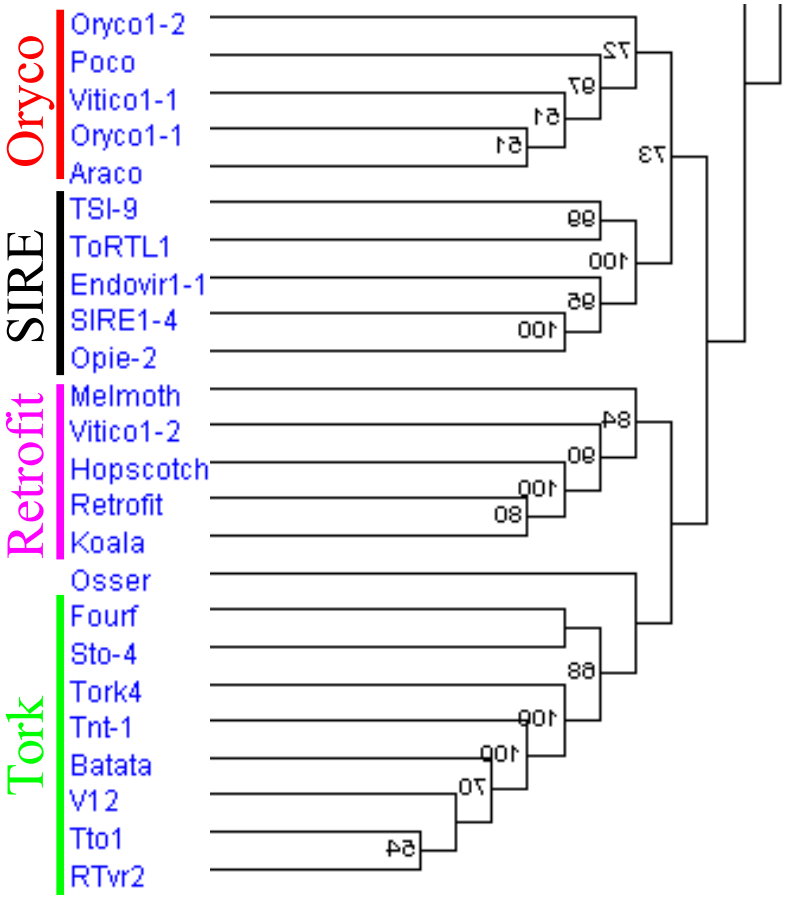


# Ty1/copia

## classification according to Whicker et al. (2007) and Llorens et al. (2011)



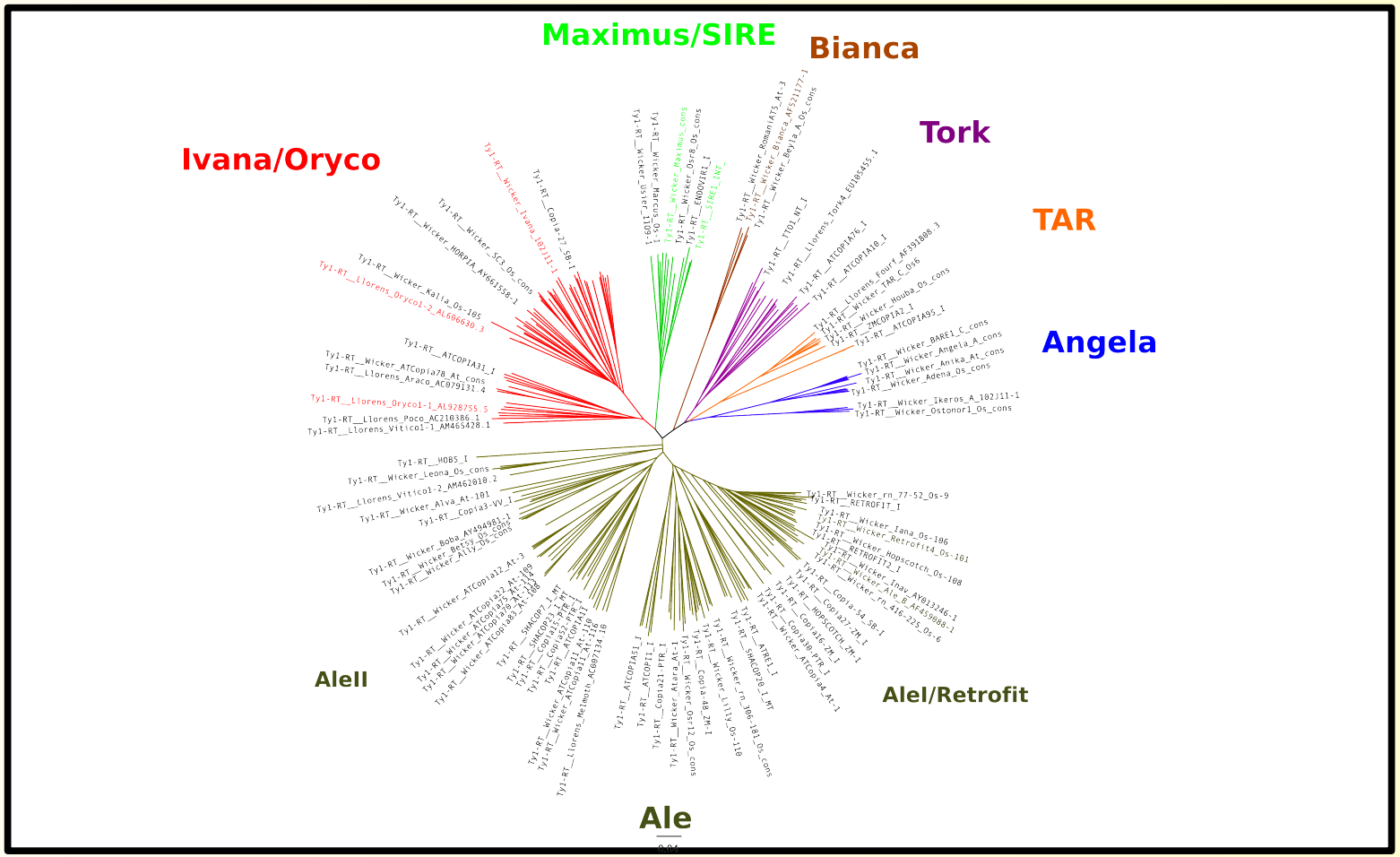

Whicker et al.




Llorens et al.

# Ty1/copia

## modified classification according to Whicker et al. (2007) and Llorens et al. (2011)



# A few tips at the end

- bad alignment leads to a wrong tree
- wrong translation leads to a wrong result
  - continuous translations from multiple frames is convenient but not always correct; consider avoiding old severely mutated sequences - typically but not necessarily those in clusters composed of only a few reads
- analyze Ty3/gypsy and Ty1/copia separately (they are too divergent and difficult to align)
- analyze more domains to get the highest confidence
  - RT, RNaseH, and INT are easy to compare
  - GAG, PROT, CHDII/CHDCR are highly divergent and often difficult to compare but can be used for finer classification of LTR retrotransposons
- if you are not sure take a look at other features (pbs, introns, extra ORF)
- it should be **you** who makes the final decision; do not blindly rely on the automatic outputs

# Coming soon ...

- DNA transposons: Full-length sequences of transposases will be added to the database (the short fragments will be removed)
- Protein sequences of extra ORFs present in some TEs will be added to the database (Athila, OGRE, PMB domain)
- Classification table will contain all types of TEs (now there are only LTR retrotransposons, other types are recognizable by codes in domains names)
- LINEs will be classified into lineages and clades (L1 – Tal1, BNR, Cin4 .., RTE - ?)
- Other improvements based on your feedback ...