# Anotação de elementos transponíveis: princípios gerais e comparação entre metodologias

Drª Desirrê Petters-Vandresen

# Por que anotar elementos transponíveis (TEs)?

- Porção significativa do conteúdo de muitos genomas eucarióticos
- Associados com aumento no tamanho de um genoma sem aumento do conteúdo gênico
- Fontes de variabilidade genética e frequentemente associados à genes de patogenicidade, conflitos e interação com hospedeiros

# Classificação de TEs (Classe I)

- Mecanismo de "**cópia e cola**" utilizando transcriptase reversa, com **RNA** como molécula intermediária da transposição

- **<u>Autônomos:</u>** transcriptase reversa funcional

- **<u>Não-autônomos:</u>** dependem de transcriptases reversas externas

- Capazes de aumentar o número total de cópias dentro do genoma

| Classification | | Structure | TSD | Code | Occurrence |
|---|---|---|---|---|---|
| Order | Superfamily | | | | |
| *Class I (retrotransposons)* | | | | | |
| LTR | *Copia* | GAG AP INT RT RH | 4–6 | RLC | P, M, F, O |
| | *Gypsy* | GAG AP RT RH INT | 4–6 | RLG | P, M, F, O |
| | *Bel–Pao* | GAG AP RT RH INT | 4–6 | RLB | M |
| | *Retrovirus* | GAG AP RT RH INT ENV | 4–6 | RLR | M |
| | *ERV* | GAG AP RT RH INT ENV | 4–6 | RLE | M |
| DIRS | *DIRS* | GAG AP RT RH YR | 0 | RYD | P, M, F, O |
| | *Ngaro* | GAG AP RT RH YR | 0 | RYN | M, F |
| | *VIPER* | GAG AP RT RH YR | 0 | RYV | O |
| PLE | *Penelope* | RT EN | Variable | RPP | P, M, F, O |
| LINE | *R2* | RT EN | Variable | RIR | M |
| | *RTE* | APE RT | Variable | RIT | M |
| | *Jockey* | ORF1 APE RT | Variable | RIJ | M |
| | *L1* | ORF1 APE RT | Variable | RIL | P, M, F, O |
| | *I* | ORF1 APE RT RH | Variable | RII | P, M, F |
| SINE | *tRNA* | | Variable | RST | P, M, F |
| | *7SL* | | Variable | RSL | P, M, F |
| | *5S* | | Variable | RSS | M, O |

**Structural features**

Long terminal repeats; Terminal inverted repeats; Coding region; Non-coding region; Diagnostic feature in non-coding region; Region that can contain one or more additional ORFs

**Protein coding domains**

AP, Aspartic proteinase; APE, Apurinic endonuclease; ATP, Packaging ATPase; C-INT, C-integrase; CYP, Cysteine protease; EN, Endonuclease; ENV, Envelope protein; GAG, Capsid protein; HEL, Helicase; INT, Integrase; ORF, Open reading frame of unknown function; POL B, DNA polymerase B; RH, RNase H; RPA, Replication protein A (found only in plants); RT, Reverse transcriptase; Tase, Transposase (* with DDE motif); YR, Tyrosine recombinase; Y2, YR with YY motif

**Species groups**

P, Plants; M, Metazoans; F, Fungi; O, Others

Adaptado de:
WICKER et al. 2007. **Nature Reviews Genetics**. DOI: 10.1038/nrg2165

# Classificação de TEs (Classe II)

- Mecanismo de "**corta e cola**" utilizando tranposase, com **DNA** como molécula intermediária da transposição

- **<u>Autônomos:</u>** transposase funcional

- **<u>Não-autônomos:</u>** dependem de tranposases externas

- Aumentam o número total de cópias dentro do genoma em condições restritas (ex: durante a replicação ou reparo de DNA)

| | | | | | |
|---|---|---|---|---|---|
| *Class II (DNA transposons) - Subclass 1* | | | | | |
| TIR | *Tc1–Mariner* | Tase* | TA | DTT | P, M, F, O |
| | *hAT* | Tase* | 8 | DTA | P, M, F, O |
| | *Mutator* | Tase* | 9–11 | DTM | P, M, F, O |
| | *Merlin* | Tase* | 8–9 | DTE | M, O |
| | *Transib* | Tase* | 5 | DTR | M, F |
| | *P* | Tase | 8 | DTP | P, M |
| | *PiggyBac* | Tase | TTAA | DTB | M, O |
| | *PIF–Harbinger* | Tase* ORF2 | 3 | DTH | P, M, F, O |
| | *CACTA* | Tase ORF2 | 2–3 | DTC | P, M, F |
| Crypton | *Crypton* | YR | 0 | DYC | F |
| *Class II (DNA transposons) - Subclass 2* | | | | | |
| Helitron | *Helitron* | RPA Y2 HEL | 0 | DHH | P, M, F |
| Maverick | *Maverick* | C-INT ATP CYP POL B | 6 | DMM | M, F, O |

**Structural features**

Long terminal repeats; Terminal inverted repeats; Coding region; Non-coding region; Diagnostic feature in non-coding region; Region that can contain one or more additional ORFs

**Protein coding domains**

AP, Aspartic proteinase; APE, Apurinic endonuclease; ATP, Packaging ATPase; C-INT, C-integrase; CYP, Cysteine protease; EN, Endonuclease; ENV, Envelope protein; GAG, Capsid protein; HEL, Helicase; INT, Integrase; ORF, Open reading frame of unknown function; POL B, DNA polymerase B; RH, RNase H; RPA, Replication protein A (found only in plants); RT, Reverse transcriptase; Tase, Transposase (* with DDE motif); YR, Tyrosine recombinase; Y2, YR with YY motif

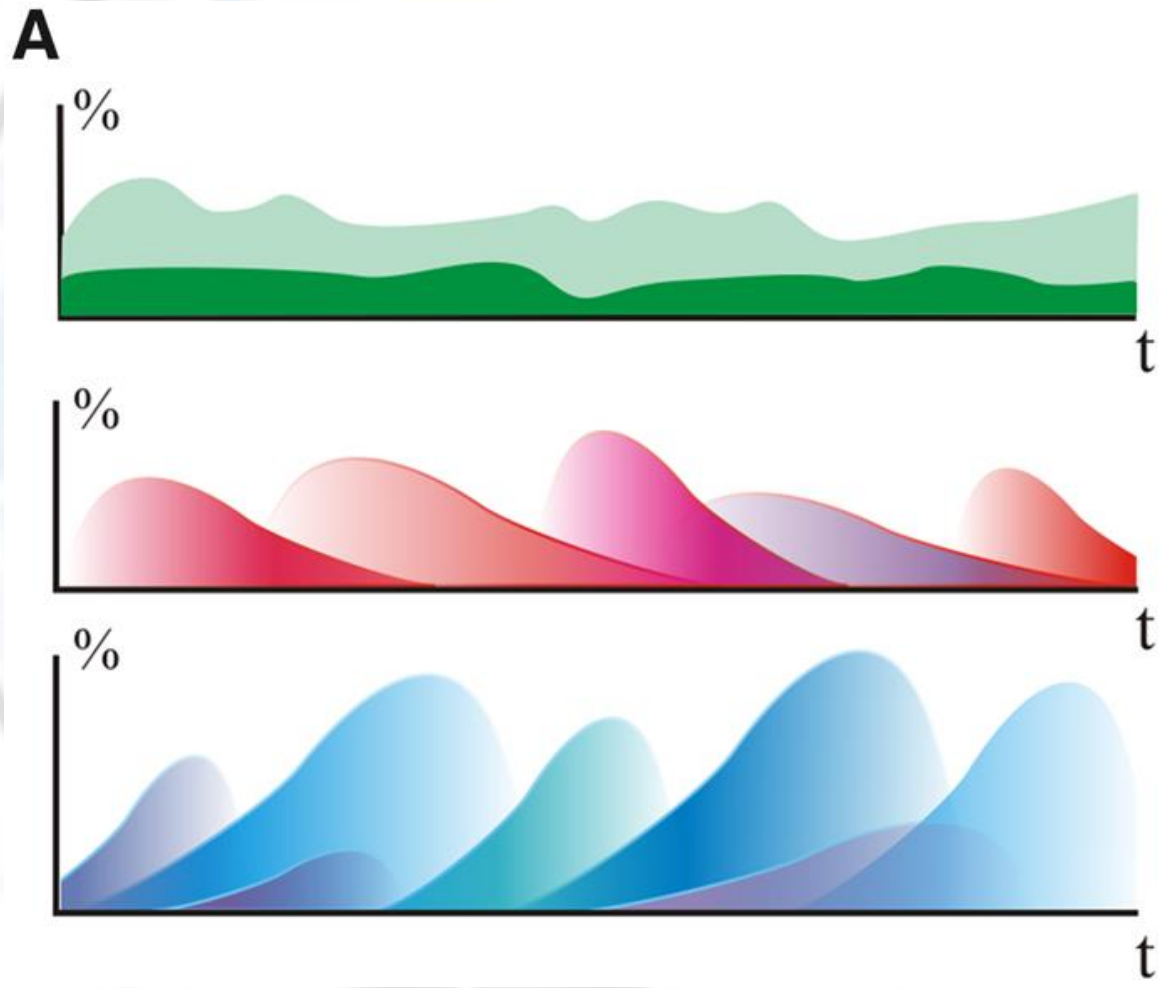
**Species groups**

P, Plants; M, Metazoans; F, Fungi; O, Others

Adaptado de:
WICKER et al. 2007. **Nature Reviews Genetics**. DOI: 10.1038/nrg2165

# Dinâmica de TEs dentro de genomas

- **<u>Verde:</u>** TEs “benignos”, neutros ao hospedeiro, com equilíbrio no número de cópias
- **Vermelho:** TEs “agressivos” que invadem o genoma, se expandem e aumentam o número de cópias. O sistema de defesa do hospedeiro não é tão eficiente para reconhecer os elementos invasores e o decaimento no número de cópias é lento
- **Azul:** TEs “dormentes” que possuem picos de expansão dentro do genoma hospedeiro e aumentam o número de cópias. O sistema de defesa é eficiente no reconhecimento das cópias e o decaimento é rápido.




Adaptado de:
ARKHIPOVA 2018. **Molecular Biology and Evolution**. DOI: 10.1093/molbev/msy083

# Reconhecimento e anotação de TEs

- **Desafio:** detectar todos as repetições e TEs, inclusive cópias degeneradas e silenciadas, e que podem não apresentar todas as características de uma cópia ativa

- Estratégias gerais:
  - *De novo*
  - Baseada em homologia
  - Baseada em repetitividade

# *De novo* – Visão geral

- Diretamente baseada na sequência genômica analisada
- Limitada à elementos que ainda apresentem as características de reconhecimento ao longo da sequência, sem degeneração
- Modelos estatísticos treinados para encontrar características presentes em nos TEs de Classe I e Classe II de diferentes ordens e superfamílias
  - Presença de repetições terminais invertidas
  - Presença de duplicação do sítio alvo
  - Presença de ORFs
  - Presença de sequências associadas a transcriptases reversas, transposases...

| Classification | | Structure | TSD | Code | Occurrence |
|---|---|---|---|---|---|
| Order | Superfamily | | | | |
| *Class I (retrotransposons)* | | | | | |
| LTR | *Copia* | GAG AP INT RT RH | 4–6 | RLC | P, M, F, O |
| | *Gypsy* | GAG AP RT RH INT | 4–6 | RLG | P, M, F, O |
| | *Bel–Pao* | GAG AP RT RH INT | 4–6 | RLB | M |
| | *Retrovirus* | GAG AP RT RH INT ENV | 4–6 | RLR | M |
| | *ERV* | GAG AP RT RH INT ENV | 4–6 | RLE | M |
| DIRS | *DIRS* | GAG AP RT RH YR | 0 | RYD | P, M, F, O |
| | *Ngaro* | GAG AP RT RH YR | 0 | RYN | M, F |
| | *VIPER* | GAG AP RT RH YR | 0 | RYV | O |
| PLE | *Penelope* | RT EN | Variable | RPP | P, M, F, O |

Adaptado de:
WICKER et al. 2007. **Nature Reviews Genetics**. DOI: 10.1038/nrg2165

# Baseada em homologia

- Métodos que usam homologia para encontrar os TEs
  - Comparação com sequências de TEs previamente caracterizados, disponíveis em bases de dados (ex: Repbase, Dfam) por meio de alinhamentos
  - Comparação com bibliotecas de elementos de vários grupos de organismos, ou restritas ao grupo do organismo analisado

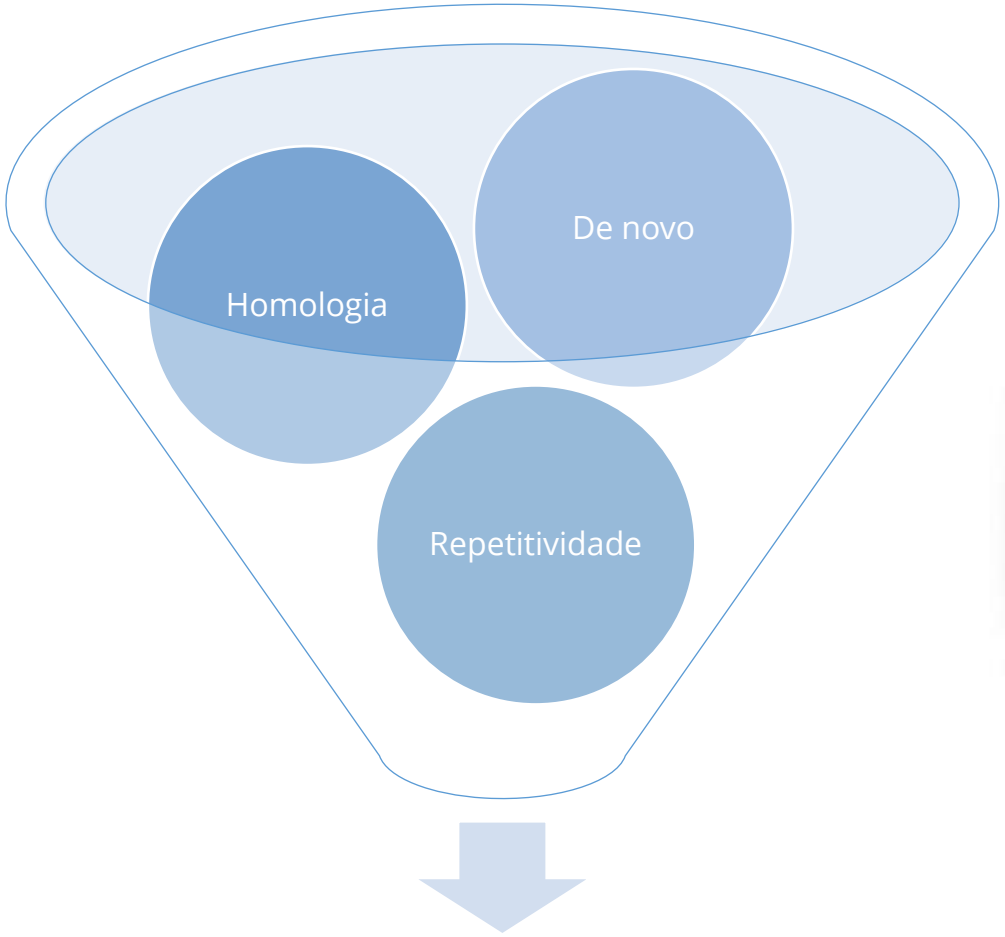

- **Limitação:** restrita aos TEs já estudados, e TEs muito degenerados podem não ser detectados

# Baseada em repetitividade

- Detecção de sequências ou blocos de sequências repetidas em comparação com o restante do genoma

- Útil para detecção de sequências teloméricas, centroméricas, DNA satélite ou sequências repetitivas que não compõem elementos transponíveis diretamente

# Abordagem híbrida: consenso e re-anotação




- Maior eficiência em comparação ao uso de abordagens isoladas
- Maior sensibilidade em comparação ao uso de bases de dados abrangentes
- Computacionalmente mais exigente: elaboração de pipelines com vários softwares e uso de servidores

Consenso e base inicial focada no genoma analisado → Nova avaliação baseada em homologia → Consenso final de todas as rodadas de avaliação, classificação e identificação

# Arquivo de saída no formato GFF3

- Estrutura similar ao arquivo de anotação gênica
- Coluna 9 normalmente contém informações sobre o elemento detectado, classificação e identificação
- Exemplos:
  - **Linha 04:** Elemento não categorizado
  - **Linha 05:** LTR-Gypsy (Classe I) incompleto
  - **Linha 06:** LINE (Classe I) incompleto

```
4  scaffold_10 Cap173_annot_REPET_TEs  match   396970  397129  0.0 +   .
   ID=ms387_scaffold_10_noCat_Cap173-B-G1-Map20;Target=noCat_Cap173-B-G1-Map20 1772
   1931;TargetLength=2500;TargetDescription=CI:NA struct:(SSR: (TAAA)6_end SSRCoverage:0.22);Identity=65.6
5  scaffold_10 Cap173_annot_REPET_TEs  match   387699  388154  0.0 -   .
   ID=ms388_scaffold_10_RLX-incomp_Cap173-B-G19-Map4;Target=RLX-incomp_Cap173-B-G19-Map4 1
   457;TargetLength=5411;TargetDescription=CI:35 coding:(TE_BLRtx: Gypsy-3-I_AF:ClassI:LTR:Gypsy: 5.29% |
   Gypsy-34_BG-I:ClassI:LTR:Gypsy: 9.46% TE_BLRx: Gypsy-33_BG-I_1p:ClassI:LTR:Gypsy: 56.79% |
   Gypsy-9_BG-I_3p:ClassI:LTR:Gypsy: 6.75% profiles: _RT_pyggy_NA_RT_NA: 99.56%(99.56%) |
   _INT_crm_NA_INT_NA: 31.50%(31.50%) | _RNaseH_maggy_NA_RH_NA: 89.26%(89.26%)) struct:(TElength: >4000bps)
   other:(SSRCoverage:0.38);Identity=83.4
6  scaffold_10 Cap173_annot_REPET_TEs  match   394840  395221  0.0 -   .
   ID=ms390_scaffold_10_RIX-incomp_Cap173-B-G3-Map20;Target=RIX-incomp_Cap173-B-G3-Map20 1399
   1755;TargetLength=2003;TargetDescription=CI:27 coding:(TE_BLRx: Tad1-14_BG_2p:ClassI:LINE:I: 12.36% |
   Tad1-31B_BG_2p:ClassI:LINE:I: 8.76%) struct:(TElength: >1000bps) other:(SSRCoverage:0.44);Identity=68.4
```