REPET: pipelines for the identification and annotation of transposable elements in genomic sequences

Timothée Flutre

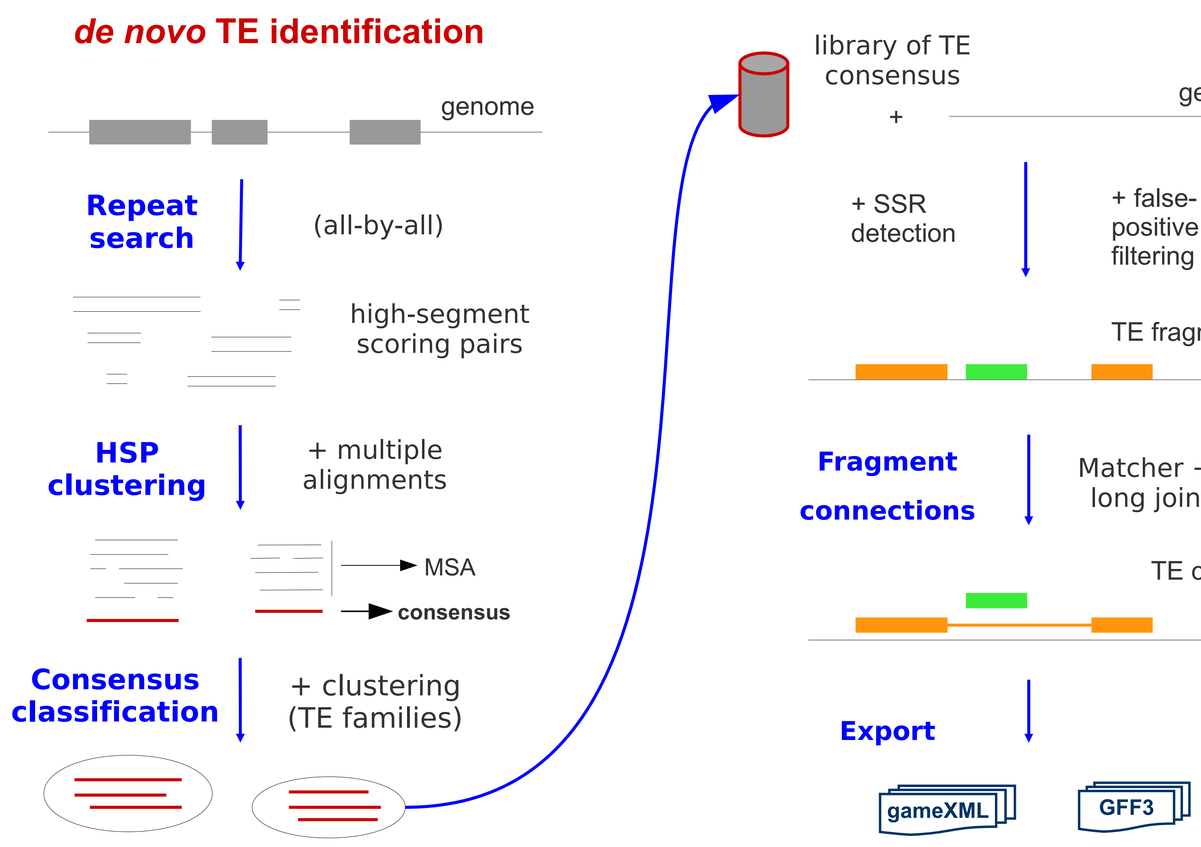
Timothée Flutre[1], Olivier Inizan[1], Claire Hoede[1], Hadi Quesneville[1]

[1]Unité de Recherche en Génomique-Info UR 1164, INRA, route de Saint Cyr, 78026 Versailles, France.

http://urgi.versailles.inra.fr/development/repet/ - urgi-repet@versailles.inra.fr

Transposable elements (TEs) are repeated genomic sequences almost ubiquitous among prokaryote and eukaryote genomes. They are acknowledged as main agents involved in genome structure dynamics but can also be viewed as “controlling” elements involved in epigenetics mechanisms and the tinkering of regulatory networks. As the number of sequencing projects is ever increasing, from model species to less studied ones, efficient approaches are required to overcome the challenge of detecting nested, fragmented TEs in large, newly sequenced genomes. In this aim we implemented a combined *de novo* pipeline, TEdenovo (Flutre *et al.* in preparation), now part of the REPET package along with the already-existing annotation pipeline, TEannot (Quesneville *et al.* 2005). These tools allow not only to detect and annotate the TE content of any sequenced genome but also to highlight the diversity of TE families by quantifying their structural variants.

I. Computational approach (REPET)



II. Genomic sequences and TE content

Drosophila melanogaster release CAF1 (130 Mb)
- all-by-all coverage: 7.4%
- *de novo* consensus: 782

Arabidopsis thaliana release TAIR9 (120 Mb)
- all-by-all coverage: 13.5%
- *de novo* consensus: 1504

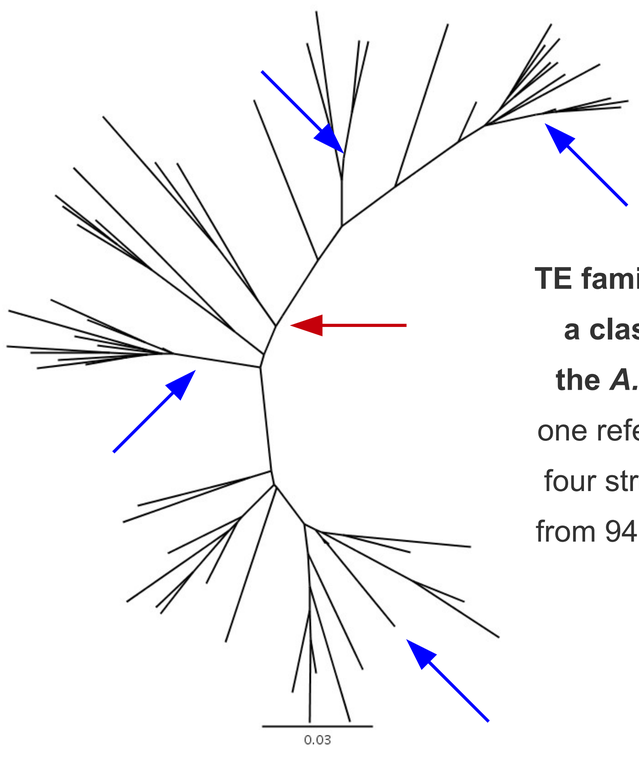
Oryza sativa release MSU6 (372 Mb)
- all-by-all coverage: 33.57%
- *de novo* consensus: 11054

Brachypodium distachyon release 1.0 (271 Mb)
- all-by-all coverage: 30.15%
- *de novo* consensus: 6741

Arabidopsis lyrata release release 1.0 (207 Mb)
- all-by-all coverage: 32.65%
- *de novo* consensus: 5229

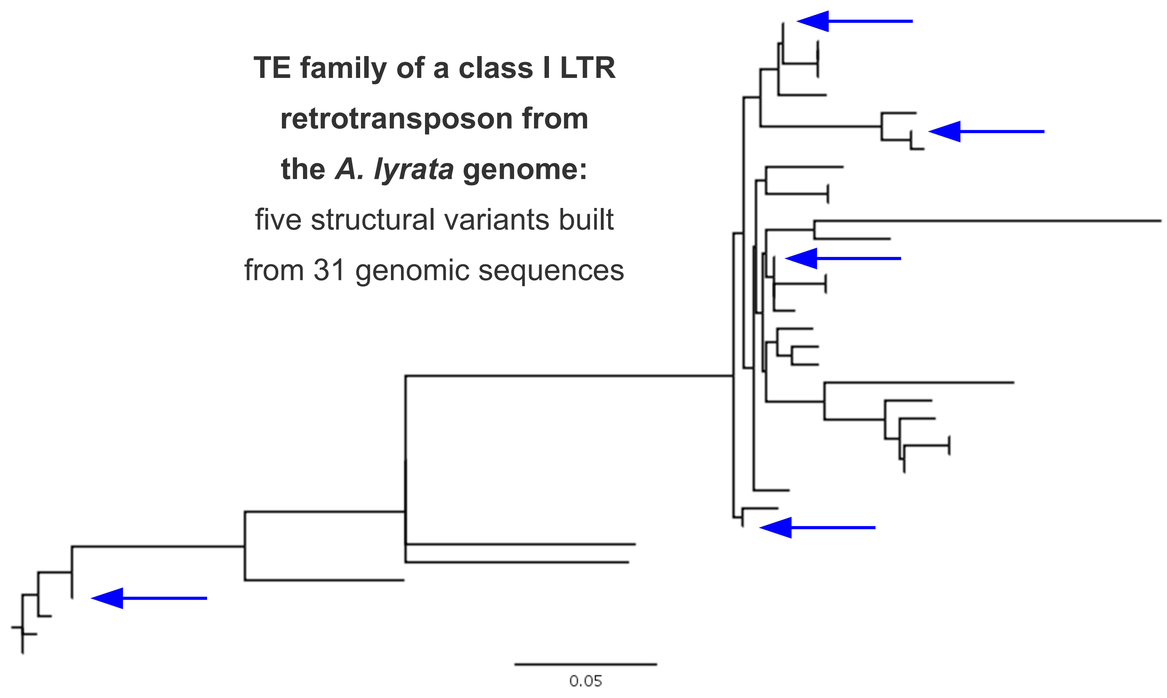
III. Diversification of TE families and emergence of structural variants

For each TE family, we built a multiple alignment (program *refalign*) and its phylogeny (program *phyml*) with the reference element (if any), the *de novo* consensus and the genomic sequences (≥3) from which each consensus was derived. Trees plotted with FigTree.



TE family of « ATREP2A », a class II Helitron from the *A. thaliana* genome: one reference sequence but four structural variants built from 94 genomic sequences

TE family of a class I LTR retrotransposon from the *A. lyrata* genome: five structural variants built from 31 genomic sequences



IV. Conclusions and perspectives

- A given TE family is best represented by several *de novo* consensus, each one corresponding to a specific structural variant.
- Our tools allow to detect such structural variants automatically and thus to quantify the degree of diversification per TE family.
- We are improving the TE classification (HMM profiles) and the definition of TE families using phylogenies of TE copies and consensus.
- We are currently analyzing TE dynamics in their « genomic ecosystem » and studying their impacts on the evolution of genome size.

Acknowledgments: I thank all the members of the URGI, especially Emmanuelle Permal, Joëlle Amselem and Victoria Dominguez for helping me in this analysis, as well as the members of the development team and the platform administrators.