Exploring the diversification of transposable elements with *de novo* approaches in whole genome sequences

Timothée Flutre

Timothée Flutre[1], Elodie Duprat[2], Catherine Feuillet[3], Hadi Quesneville[1]

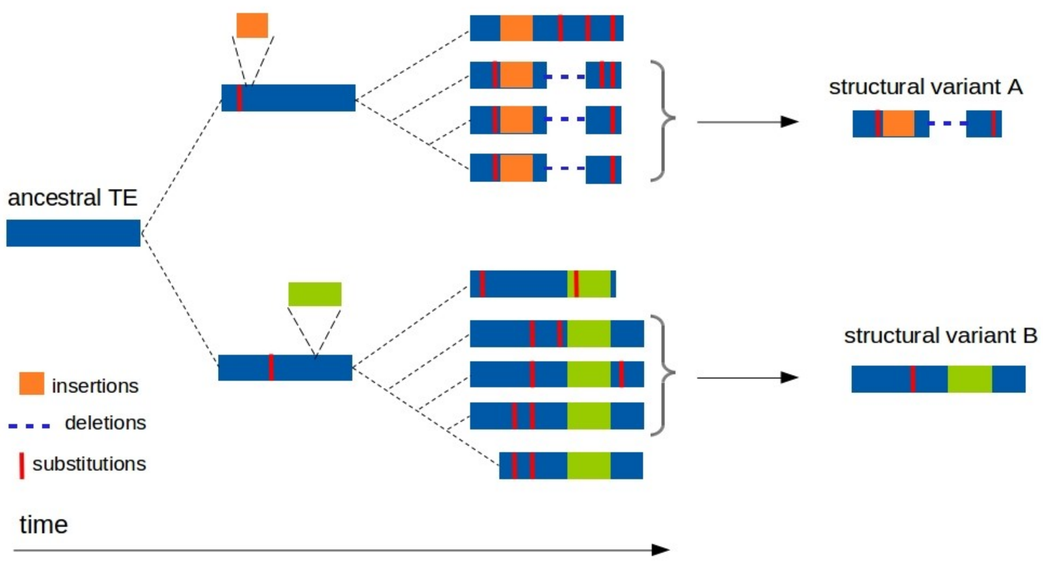
[1] Unité de Recherche en Génomique-Info UR1164, INRA de Versailles-Grignon, Route de Saint Cyr, Versailles, 78026, France
[2] Protéines: biochimie structurale et fonctionnelle, UFR 927, Université Pierre-et-Marie-Curie, 7 quai Saint Bernard, 75252 Paris cedex 5, France
[3] Génétique, Diversité et Ecophysiologie des Céréales, UMR 1095, INRA site de Crouël, 234 avenue du Brézet, 63100 Clermont-Ferrand, France

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

Transposable elements (TEs) are mobile 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.* submitted), 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, a first step in the understanding of genome ecology.

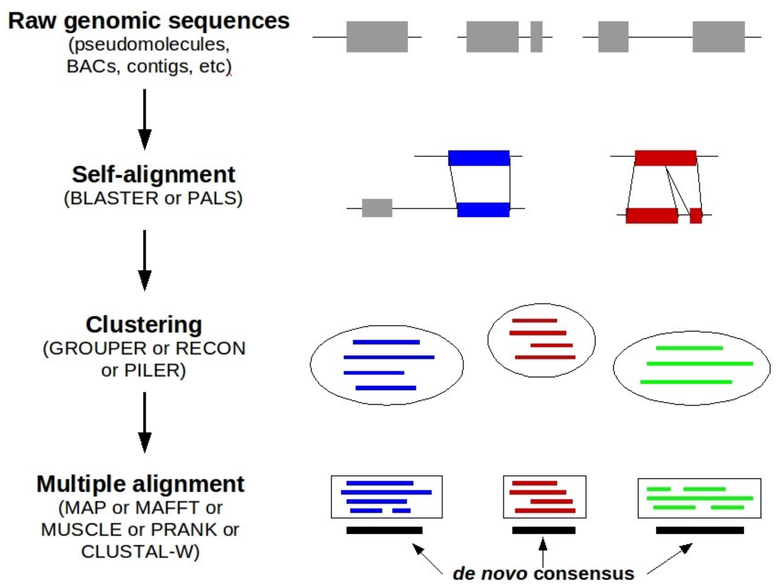
1

Aim: recovering all the ancestral TE sequences that transposed while distinguishing their structural variants that emerged throughout evolution



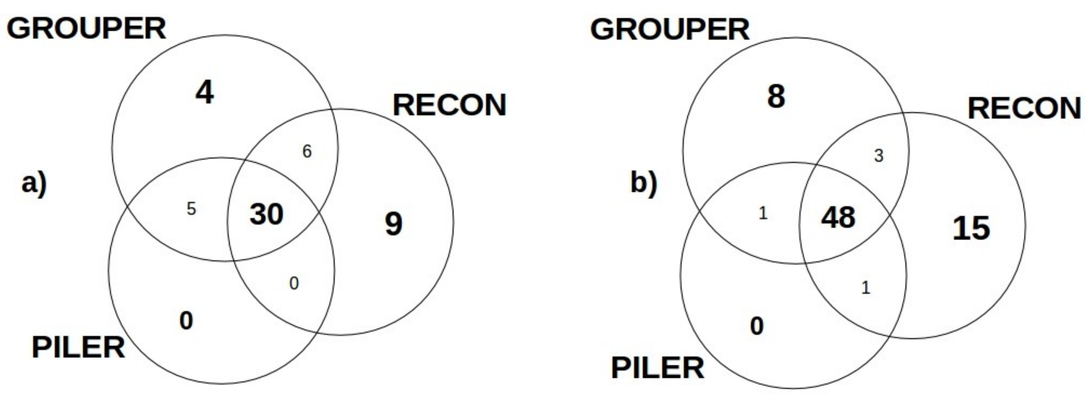
2

Comparative analysis: three first steps of the TEdenovo pipeline



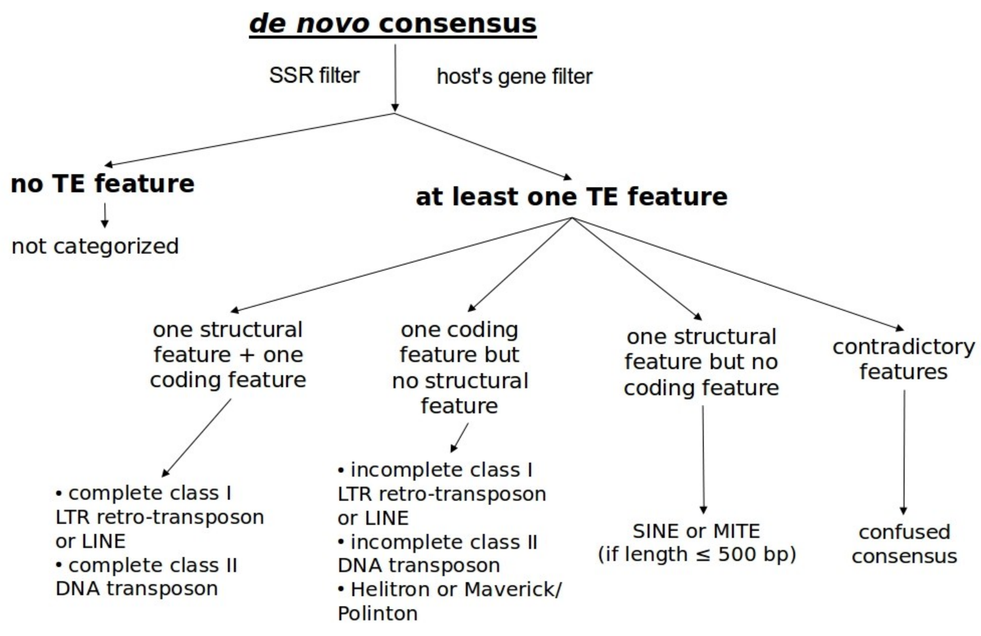
3

Venn diagram showing the gains achieved by combining several clustering programs — GROUPER and RECON in particular — in terms of the TE sequences fully recovered from the *D. melanogaster* genome (a) and from the *A. thaliana* genome (b).



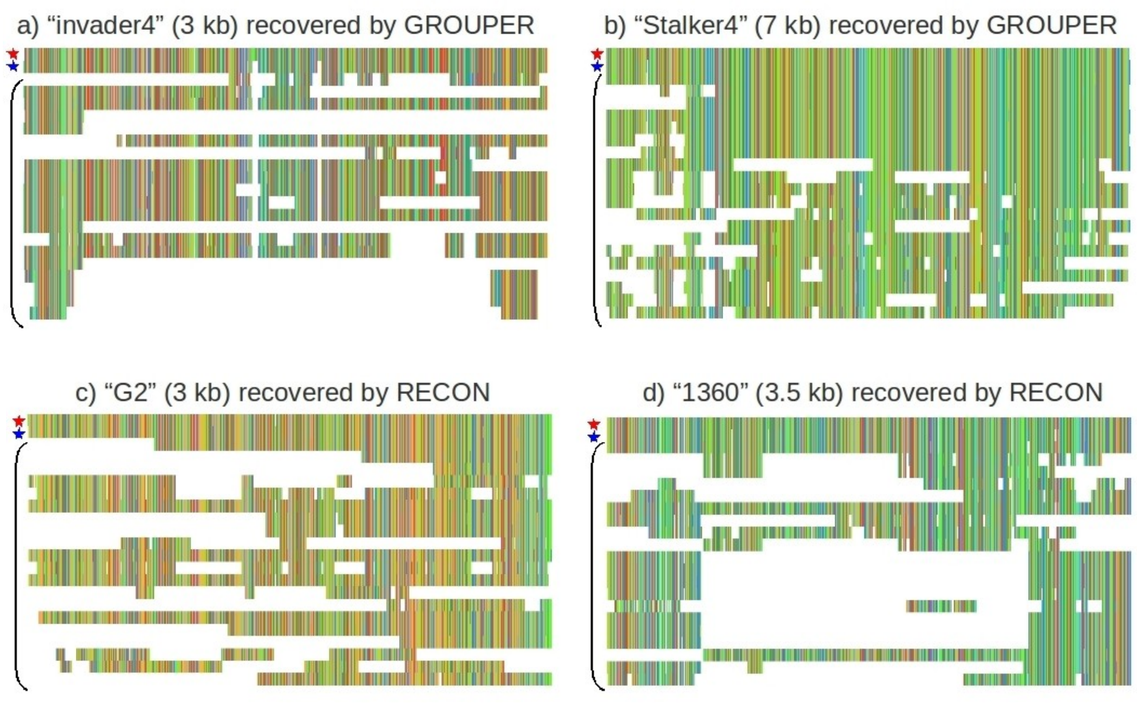
4

Simplified decision tree implemented to classified TE sequences at the end of the TEdenovo pipeline.



5

Extensive structural variations between copies of several TE families as recovered by the TEdenovo and TEannot pipelines.



6

TE annotation results obtained with reference and *de novo* databanks.

Genome	TE library	Consensus sequences	TE genome coverage	Number of copies	S_n	S_p
D. melanogaster	BDGP	125	10.51%	31208	NA	NA
	GROUPER	712	10.29%	43699	81.92%	98.12%
	RECON	437	11.05%	33072	87.77%	97.95%
	PILER	114	8.87%	32789	74.07%	98.79%
	G+R+P	568	11.98%	42847	91.43%	97.35%
A. thaliana	Repbase	318	19.02%	41146	NA	NA
	GROUPER	1237	18.78%	41791	79.29%	95.43%
	RECON	1004	23.69%	49470	88.75%	91.59%
	PILER	300	13.14%	34818	56.56%	97.05%
	G+R+P	1232	22.77%	44059	87.03%	92.32%

Conclusions and perspectives

- A given TE family can be 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 wish to thank all the members of the URGI, especially Emmanuelle Permal, Joëlle Amselem and Victoria Dominguez for for their fruitful remarks, the members of the development team Olivier Inizan and Claire Hoede, and the platform administrators Sébastien Reboux and Isabelle Luyten.