TEFLoN2, an automized and accurate computational tool for detecting and analyzing insertions of transposable elements in population data

Anna-Sophie Fiston-Lavier*[1,2], Corentin Marco[2,3], Clothilde Chenal[2,3,4], and Fontaine Michael[4,5]

[1]Institut des Sciences de l'Evolution - Montpellier – CNRS : UMR5554, Université Montpellier II - Sciences et techniques – Place E. Bataillon CC 064 34095 Montpellier Cedex 05, France
[2]Institut Universitaire de France (IUF) – IUF – Paris, France
[3]Institut des Sciences de l'Evolution - Montpellier – CNRS : UMR5554 – Place E. Bataillon CC 064 34095 Montpellier Cedex 05, France
[4](MIVEGEC-IRD, CNRS, University of Montpellier) in Montpellier, FRANCE – UMR 224 MIVEGEC (IRD, UM, CNRS), Montpellier – Montpellier, France
[5]University of Groningen – Groningen, Netherlands

Abstract

Transposable elements (TEs) are mobile, repetitive and mutagenic elements of DNA known to be important components of eukaryotic genomes and major actors in genome evolution. In order to estimate their impact on genome evolution, we need to detect them from individual to populations. Detection of TE insertions using paired-end reads revealed a high false-positive rate, up to 40%. Even if long–read sequencing technologies overcome this detection issue, this type of sequencing is not suitable for population genomics studies that require a large amount of resequencing data from multiple samples and for a reasonable cost.
One of the most promising tools highlighted in previous benchmark studies is TEFLoN that uses short-read pooled-data. While TEFLoN is easy to install and use, many technical limitations have been identified, such as the fact that each script must be launched independently without parallelization that makes it time and memory consuming. Thus, we developed an automatic and optimized version of this tool called TEFLoN2 by upgrading the code and developing a SnakeMake pipeline. We also developed a new module accurately estimating the TE frequency in pooled or large single sequencing dataset. We were then able to appreciate its accuracy in simulated and public resequencing data (https://github.com/asfistonlavie/TEFLoN2).

Keywords: Transposable elements, structural variants, annotation, snakemake pipeline, population genomics

*Speaker