New insights offered by next-generation sequencing data for transposable element discovery and annotation

Anna-Sophie Fiston-Lavier*[1]

[1]Institut des Sciences de l'Evolution - Montpellier (ISEM) – CNRS : UMR5554, Institut de recherche pour le développement [IRD] : UMR226 – Place E. Bataillon CC 064 34095 Montpellier Cedex 05, France

Résumé

Next-generation sequencing (NGS) approaches are fundamentally changing the way in which scientists undertake genome evolution studies. The most common way to understand how genomes vary and evolve is to study genome variations identifying and analyzing functionally relevant SNPs and structural variants. While a numerous computational tools have been developed for the automatic detection of SNPs, a more recent attention is being paid to structural variants such as transposable elements (TEs) as more and more studies support their role in genome evolution and adaptation.
We recently designed an integrated package called "T-lex" to automatically detect and analyze TE variants using short reads sequencing data. The T-lex package is composed of two distinct pipelines: the "T-lex de novo" pipeline that discovers and annotates novel TE insertions not annotated or absent from the reference genome and the "T-lex2 presence/absence" pipeline that is the only available software that allows routine, automatic and accurate genotyping of individual TE insertions and estimation of their population frequencies both using individual strain and pooled NGS data. Furthermore, T-lex2 also assesses the quality of the calls allowing the automatically detection of the TE insertion signatures (i.e., TSD). Such information combined with the T-lex2 analysis of the genomic environment for each TE insertion may conduct to the identification of miss-annotated TEs and provide enough information to re-annotate them (version 2.0; http://sourceforge.net/projects/tlex; Fiston-Lavier et al. 2011, 2014).
Another way to discover TEs accurately is to use the last genome sequencing technologies which allows synthetizing long reads such as the "Illumina TruqSeq synthetic long-reads" technology designed to overcome the assembly of complex genomes, enriched in repeats such TEs. Indeed, TE assembly remains challenging as they introduce ambiguity during genome reconstruction. Testing the veracity of this technology to place individual TE copies in their proper genomic locations as well as accurate reconstruction of TE sequences, we highlighted the difficulty to annotate especially TEs from families exhibiting high sequence identity, high copy number, or complex genomic arrangements (McCoy et al. 2014).

*Intervenant