sideRETRO: structural variations intercurences discovery environment for retrocopies

José Leonel Lemos Buzzo, Thiago Luiz Araujo Miller, Pedro Alexandre Favoretto Galante

USP

Abstract

The understanding of Transposable elements' mechanisms are gaining a rising focus on actual researches as key features of genomic structures and the impacts of their dynamics are directly associated with many pathological scenarios. It has been clearly shown the protagonic role exerted, for example, by somatic retrotransposon insertions in tumorigenenic cases, whether disrupting promoters of critical protein coding genes, or creating new splicing sites and isoforms, or even becoming expressed. Therefore, accurate detection methods ought to be settled down for them, methods by which whole genomic assessments of the copy number variations of these transposable elements would be feasible. However, some quantitative difficulties lies on this task concerning the ambiguity on mapping new inserts: actual aligners frequently report their reads as belonging to the parenthal gene in the reference genome. Corroborating this fact, literature shows only a few examples of bioinformatics tools available to this errand and, even these, cannot ensure their accuracy because of a lack in false positive statistical controls. So, to get around this problem, a candidate algorithm would need to (1) distinguish ambiguous read mappings from their parenthals and other fixed copies of it, based on an annotated gene list; and (2) robustly learn to discern the false positive cases using some simulation approach. Now focusing on retrocopies, which are new insertions of processed protein coding genes' mRNAs made by LINE retrotransposon machinery, we present sideRETRO, a computational bioinformatis tool for polymorphic and somatic retrocopies discovery, in a genomic landscape. Provided with an accuracy tuning simulation method suited for this task and a ambiguity solving engine based on discordant reads mappings, our tool was used to search for retrocopies on ten whole genomes sequenced from healthy individuals with more than eighty years old. It was chosen only healty individuals because of the need to compose a non tumorigenic retrocopies profile when comparing to future pathological samples. So, assessed by a previous simulation batch for false positive filtering, our algorithm reached a 0.87 accuracy rate on a 5-fold cross validation. And, when used on the ten whole genomes with high coverage (40x), it discovered a mean of seventeen retrocopies per genome, which were further identified when polymorphic or not and when cancer related or not.

Funding: CNPq