

---

# De novo identification of Transposable Elements from RNA-seq data

Sasha Darmon<sup>\*1</sup>, Vincent Lacroix<sup>1</sup>, and Arnaud Mary<sup>1</sup>

<sup>1</sup>Laboratoire de Biométrie et Biologie Evolutive - UMR 5558 – Université Claude Bernard Lyon 1, Institut National de Recherche en Informatique et en Automatique, Centre National de la Recherche Scientifique – 43 Bld du 11 Novembre 1918 69622 VILLEURBANNE CEDEX, France

## Abstract

Short-read RNA sequencing has generated extensive collections of reads. The goal of our work is the de novo identification of repetition families from RNA-seq reads. This would help to discover novel repetition families, including transposable elements, in particular for non-model species. This could also help to improve de novo transcriptome assembly.

We are specifically working with De Bruijn graphs, an efficient data structure where every transcript corresponds to a path within this graph. Our research involves characterizing complex regions that contain families of repetitions and replacing them with consensus nodes. The objective of this novel method is to operate de novo, without relying on genomic references nor repeat consensus sequences.

Preliminary results in dog and drosophila datasets have enabled us to identify regions of the De Bruijn graph that are associated with various types of repetitions. Some of these repetitions are TEs. Out of those, we expect that some correspond to full-length active families, while others are TE-derived elements associated with TE insertions within genes.

**Keywords:** De Bruijn graph, repetition family, RNA, seq

---

<sup>\*</sup>Speaker