CARNAC-LR: De novo Clustering of Gene Expressed Variants in Transcriptomic Long Reads Data Sets
Camille Marchet, Lolita Lecompte, Corinne da Silva, Corinne Cruaud, Jean-Marc Aury, Jacques Nicolas, Pierre Peterlongo

To cite this version:

HAL Id: hal-01929963
https://hal.archives-ouvertes.fr/hal-01929963
Submitted on 21 Nov 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
**Goal: de novo cluster Nanopore reads per expressed genes**

**Data:** Nanopore 1D reads from mouse transcriptome sequenced with MinION (accession number: ERP107503)

**Results:**
- State of the art does not perform well on ONT reads
- We introduce CARNAC-LR, a new clustering approach designed for long reads
- Validations on mouse transcriptome

**Algorithm overview:**
- maximize local edge density
- minimize cut size
- partition the graph

**Pipeline overview:**
From reads to clusters per expressed gene

**Results on whole mouse transcriptome:**
- Output graphical example for mouse Picp5 gene
- Performances:
  - For 1 million reads
    - wallclock 3 hours (40 threads)
    - memory: 30G

**Main achievements**
- Clusters de novo ONT reads by expressed genes
- Scales a whole mouse transcriptome
- Performs better than state of the art on ONT reads
- Validated using comparison to mapping strategy on real data

**Tool:**
github.com/kamimrcht/CARNAC-LR

**Preprint:**
biorxiv.org/content/early/2018/03/26/170035

**Contact:** camille.marchet@irisa.fr