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:
Camille Marchet, Lolita Lecompte, Corinne da Silva, Corinne Cruaud, Jean-Marc Aury, et al..

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.
CARNAC-LR:
De novo Clustering of Gene Expressed Variants in Transcriptomic Long Reads Data Sets

Camille Marchet¹, Lolita Lecompte¹, Corinne Da Silva², Corinne Crudu¹, Jean-Marc Aury¹, Jacques Nicolas¹, Pierre Peterlongo¹

¹Univ Rennes, Inria, CNRS, IRISA, F-35000 Rennes, France
²Commissariat à l’Energie Atomique (CEA), Institut de Biologie François Jacob, Genoscope, 91000, Evry, France

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:

Key ideas:
- maximize local edge density
- minimize cut size
- partition the graph

Pipeline overview:

From reads to clusters per expressed gene

Results on whole mouse transcriptome:

Performance: For 1 million reads
¬ wallclock 3 hours (40 threads)
¬ memory: 30G

Clusters purity and completeness assessed using mapping strategy (BLAT+est2genome)

Work in progress:

Goals:
- Identify alternative isoforms from CARNAC-LR’s clusters
- Propose one consensus per isoform

Key ideas:
- intra-cluster multiple sequence alignment
- detect alternative blocks (exons)
- separated block consensus computation

Tool:
github.com/kamimrcht/CARNAC-LR
Preprint:
biorxiv.org/content/early/2018/03/26/170035

Contact: camille.marchet@irisa.fr