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.
De novo Clustering of Gene Expressed
Variants in Transcriptomic
Long Reads Data Sets

Camille Marchet1, Lolita Lecompte1, Corinne Da Silva1, Corinne Cruaud1, Jean-Marc Aury1, Jacques Nicola1, Pierre Peterlongo1
1Univ Rennes, Inria, CNRS, IRISA, F-35000 Rennes, France
*Commissariat 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:
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

Acknowledgments
★ Genoscope platform
★ ASTER ANR
★ Genouest team & infrastructures

De novo Link
Louvain
Modularity
CPM
CARNAC-LR

Recall
Precision
Jaccard index
76%
<15%
<0.1
89%
<15%
<0.1
59%
79%
98%
0.79

Recall
Precision
Status
59%
59%
error
77%
99%
run
65%
98%
run

Starcodes
Tofu
SEED
CD-16K
CARNAC-LR

Recall
61%
NA
NA
NA
61%

Precision
100%
100%
100%
100%
97%

Status
run
run
run
run
run

Datasets:
- benchmark 1: community detection algorithms
- benchmark 2: sequence clustering tools

Expected clusters:

Basal gene expression
Alternative events

Families of genes expressing transcripts

pipeline overview:
From reads to clusters per expressed gene

Rna-seq:
Minimap
CARNAC-LR
Clusters

Tool:
github.com/kamimrcht/CARNAC-LR
Preprint:
biorxiv.org/content/early/2018/03/26/170035
Contact: camille.marchet@irisa.fr

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

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

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

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

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

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