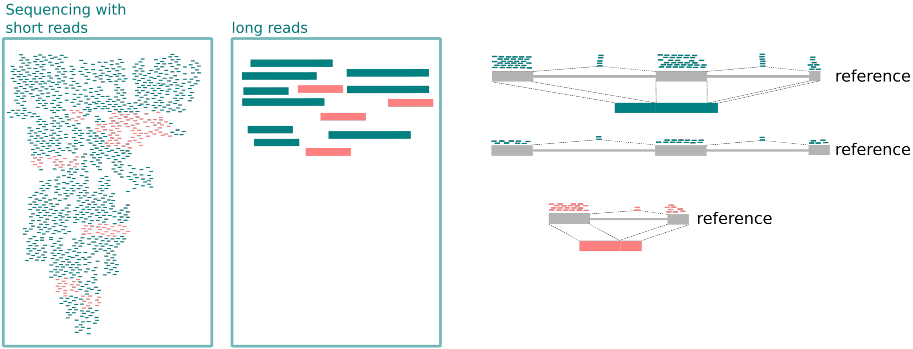
CARNAC-LR: clustering genes expressed variants from long read RNA sequencing

Camille Marchet, Lolita Lecompte,
Corinne Da Silva, Corinne Cruaud, Jean-Marc Aury,
Jacques Nicolas and Pierre Peterlongo

Workshop RNA-seq and Nanopore Sequencing – ANR ASTER

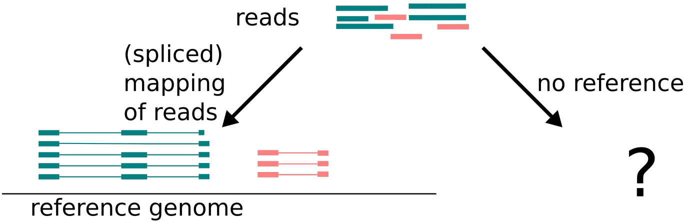
December 13th, 2017

RNA-seq and long read sequencing

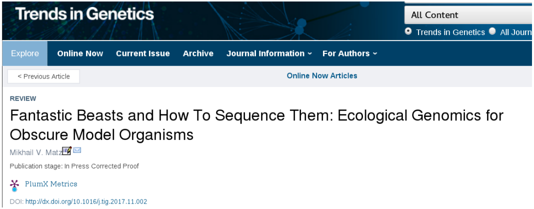


- Direct access to the different isoform structures and full-length molecules
- Avoid assembly / transcript reconstruction by mapping
- Quantification with ONT long reads [Oikonomopoulos et al. 2016]
- Annotated variants and novel variants discovery with long reads [Hoang et al. 2017, Abdhel-Ghany et al. 2016, Wang et al. 2016,...]

To map or not to map?



- Mapping of reads on reference genome (GMAP [Wu et al. 2005])
- Or transcriptome (recently Graphmap [Sovic et al. 2015])
- What if no reference ?



REVIEW

Fantastic Beasts and How To Sequence Them: Ecological Genomics for Obscure Model Organisms

Mikhail V. Matz

Publication stage: In Press Corrected Proof

PlumX Metrics

DOI: http://dx.doi.org/10.1016/j.tig.2017.11.002

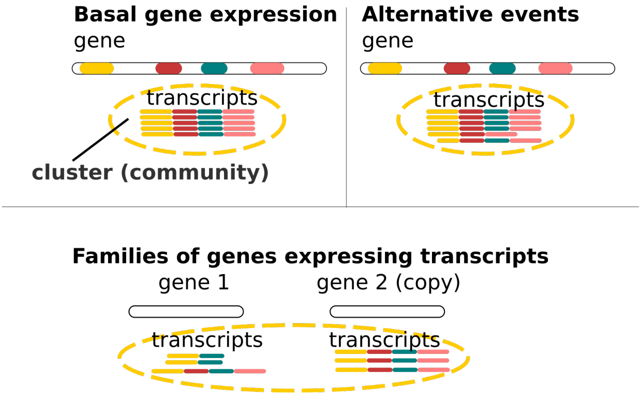
A need that starts to be expressed in the literature

- ToFu: cluster of reads by gene and isoforms detection[Gordon et al. Plos One 2015]
- Describe alternative variants: [Liu et al. Molecular ecology Resources 2017]
- Both dedicated to PacBio, need sequences of high accuracy

Our goals

- More generic approach
- Make the best of the full data set, no prior filter/treatment

Expected behavior of our clustering



Basal gene expression

gene

transcripts

cluster (community)

Alternative events

gene

transcripts

Families of genes expressing transcripts

gene 1

gene 2 (copy)

transcripts

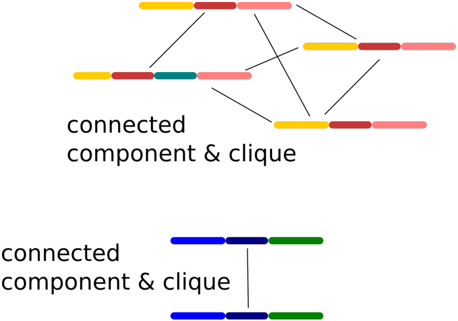
transcripts

Detect all variants for each gene de novo

Problem specificity

- Alternative variants in data
- Gene families
- Errors in reads
- Heterogeneous sizes distributions of clusters

A clustering problem: graph we work on

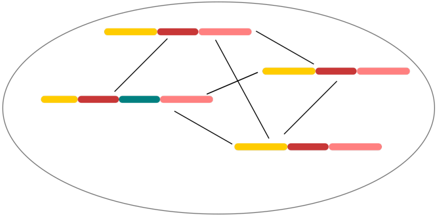


connected component & clique

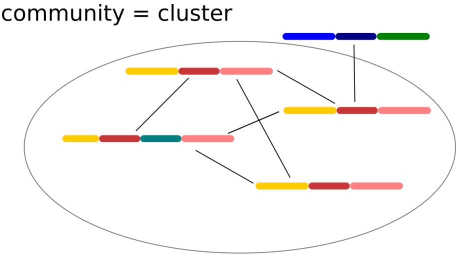
connected component & clique

A clustering problem: clusters as genes

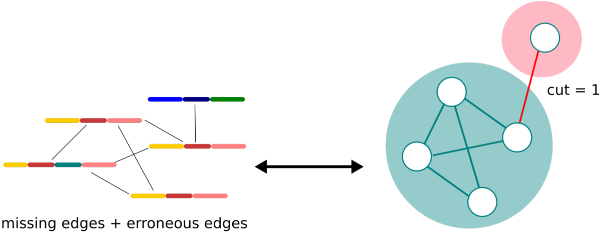
community = cluster



A clustering problem: graph in practice



A clustering problem: community detection

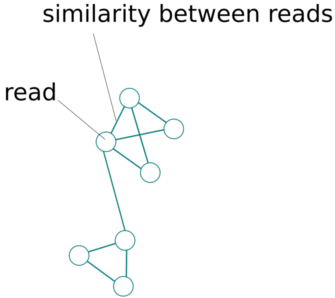


cut = 1

missing edges + erroneous edges
=> delineate a gene in the graph

- well interconnected subgraphs
- minimal cut
- disjoint sets

Detect all variants for each gene de novo

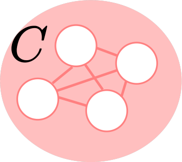


Community detection

- **Deal with the indel specificity:** detect overlaps between erroneous reads (Minimap[Li 2016], GraphMap[Sovic et al. 2015], BLASR[Chaisson et al. 2012]...)
- **Start for clustering of variants**: graph of similarity of reads

Measure of connectivity in the graph

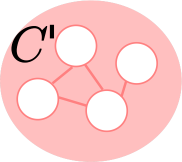
We rely on the clustering coefficient (*ClCo*) [Watts and Strogatz 1998]



Number of edges if C were a clique:6

ClCo = 6 / 6 = 1

Actual number of edges in C :6



Number of edges if C' were a clique:6

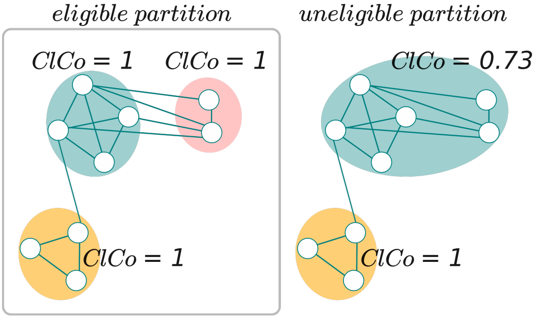
ClCo = 2 / 3

Actual number of edges in C' :4

Clustering problem

- Prop.1: A community is a connected component having a clustering coefficient above or equal to a fixed cutoff θ.
- Prop.2: Communities are disjoined sets.

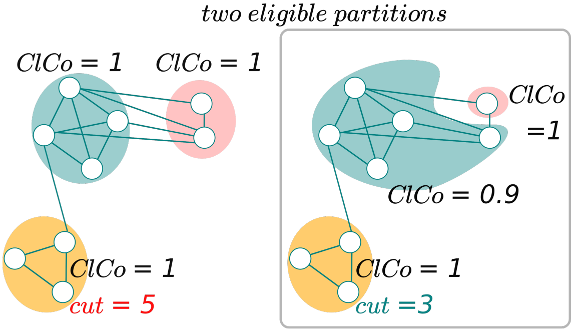
$\theta = 0.9$



Clustering problem

- Prop.3: An optimal clustering in k communities is a minimal k-cut of the graph

$$\theta = 0.9$$



- min k-cut NP hard for $k \geq 3$ [Dahlhaus et al. 1994])

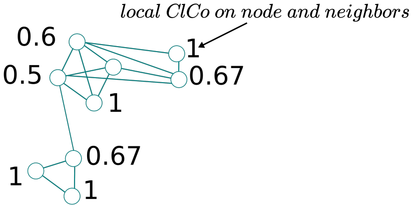
Difficulties arising from this problem

- We don't know the number of community in advance, *k*-cut NP-hard for $k \geq 3$ [?]
- The cutoff θ is not known either
- Potentially many θ values to test

Implementation: choose theta interval

- The cutoff θ is not known: test different values
- Do not compute all possible θ for all connected components

$\theta \in$[?,?]

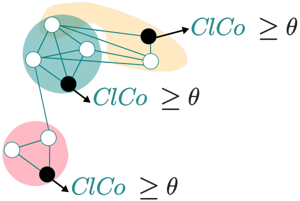


$\theta \in$\{0.5,0.6, 0.67,1\}

- Adaptive values for each connected component
- Key for scaling

Implementation: find k

1.*relax the disjoined subgraphs condition*



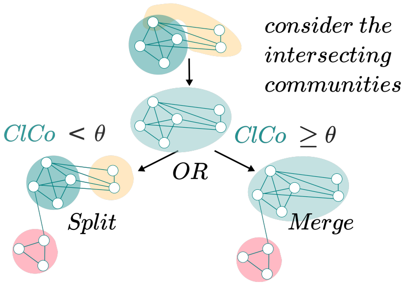
$ClCo \geq \theta$

$ClCo \geq \theta$

$ClCo \geq \theta$

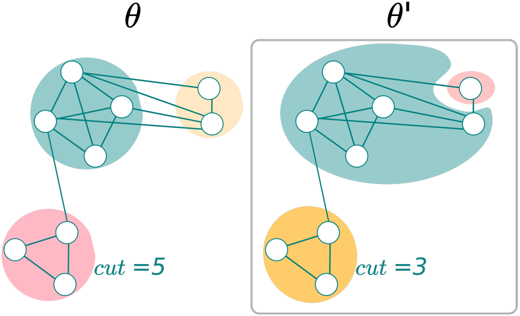
Implementation: find k

2.$refine\ the\ boudaries\ to\ obtain\ a\ partition$:



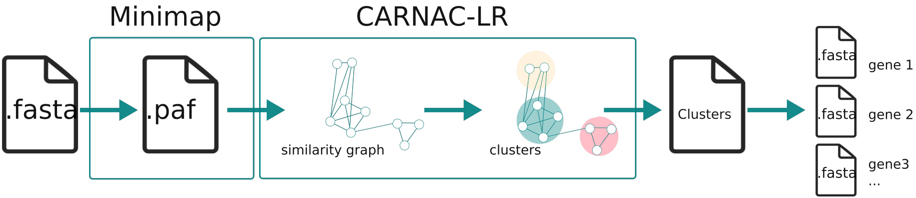
Final communities

$$different\ \theta\ values\ yield\ different\ cut\ values$$



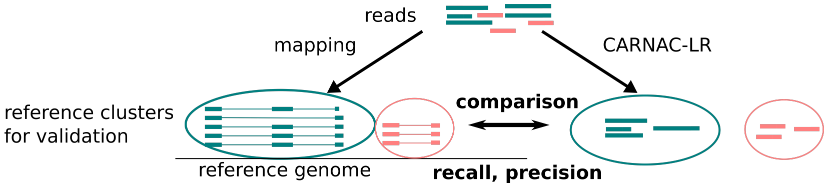
- Keep the partition associated to the minimal cut

Pipeline



github.com/kamimrcht/CARNAC

How to validate ?



- Data: **mouse transcriptome 1D Nanopore reads transcriptome**
- NB: mapping has its own limitations

Comparison to other community detection approaches

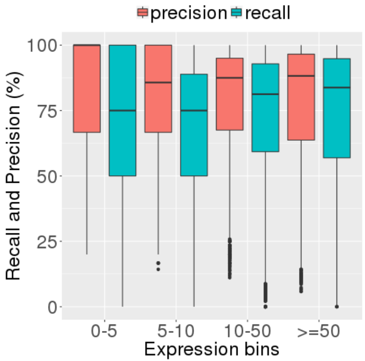
- Comparison to classic approaches: hierarchical, modularity based, CPM

CARNAC-LR pros

- Best precision
- Best trade-off between precision and recall
- Best similarity to ground truth clusters (Jaccard Index)
- No need of parameters

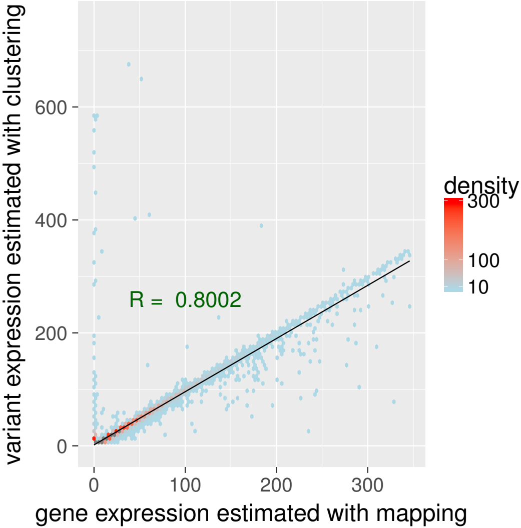
Well-tailored clustering for transcriptomic long reads

Validation real size data set



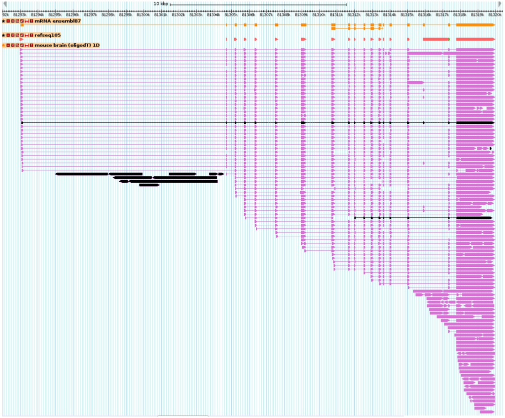
- ∼ 1M reads
- **Recall and precision not much impacted by expression levels**
- Minimap + CARNAC-LR: 3 hours using 10 threads / Mapping approach: ∼ 15 days

Proxy to genes' expression



Straightforward use of our method

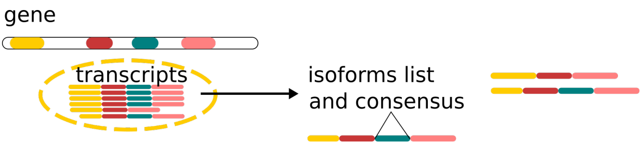
A visual example of CARNAC's output

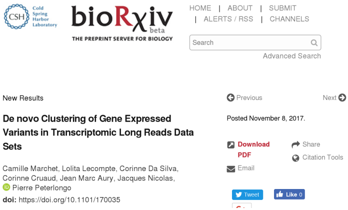


- 112 reads from a cluster output by CARNAC (purple)
- All reads map to the same locus: gene Pip5k1c (chr 10)
- 8 reads present in the data missing in the cluster (black)

Future work

- Correct by clusters and find isoforms within clusters





New Results

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

doi: https://doi.org/10.1101/170035

Posted November 8, 2017.

Conclusion

Take-home messages

- Accurate tool that outputs clusters of transcripts by gene
- Generic, first tool to perform on ONT
- For model and non model species
- Availability: `github.com/kamimrcht/CARNAC`
- Preprint

Perspectives

- Scale to meta-transcriptomics

Acknowledgments

- Dyliss, GenScale teams and Genouest platform
- Genoscope and ANR ASTER