Efficient data structure for indexing and similarity computation of nucleic sequences

(SHORT) READ CONNECTOR LINKER

Application examples in transcriptomics

C Marchet[1], A Meng[2], L Lecompte[1], A Limasset[1], L Bittner[2], P Peterlongo[1]
[1]Inria Rennes - Bretagne Atlantique/IRISA, EPI GenScale, Rennes 35042, France
[2]Sorbonne Universités, UPMC Univ Paris 06, Univ Antilles Guyane, Univ Nice Sophia Antipolis, CNRS, Evolution Paris Seine - Institut de Biologie Paris Seine (EPS - IBPS), 75005 Paris, France

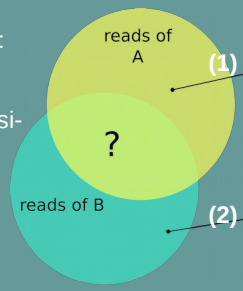
A lightweight data structure

SRCL: compute similarity between reads from sets A and B

(1) **index k-mers** of A in quasi-dictionary:
kmer → origin (read)

(2) **query k-mers** of reads from B in quasi-dictionary

(3) if a read of B shares enough k-mers with a read of A
→ **similar reads**



Quasi-dictionary

Efficient dispatching of information related to k-mers in memory + fingerprint associated

Presence/absence check for a k-mer + access to information

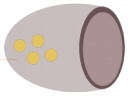
★**Probabilistic**

★**Up to billion elements indexed**

★**Fast construc-tion and query**

Marine ecology using short reads

A scaling approach to study Holobionts



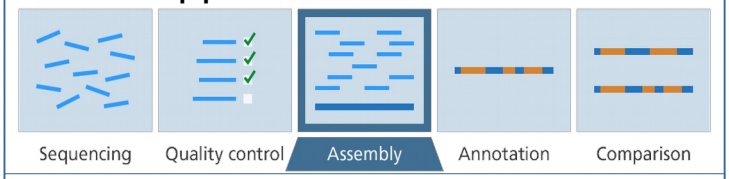
Marine holobiont:
Radiolaria + dinoflagellate
(host) (symbiont)

Can we de novo assemble clean, unambiguous sequences for the host and for the symbionts ?

Sequencing:

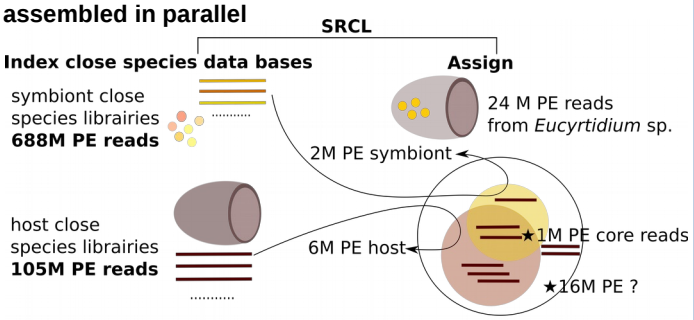
35M PE meta transcriptomics short reads
No reference
Mix of reads from host and symbiont

Global pipeline:



Contribution:

SRCL helps segregating reads from the holobiont into 2 **distincts sets** (a host and a symbion) which are then de **novo assembled in parallel**



Tackle assembly difficulties:
-High volumes indexed and processed
-Novel transcripts assembly
-Validation to come

→ **Easy and highly scalable index and query for large metaomics data sets**

Facing long reads (PB/ONT) challenges

A long reads proof of concept tool

- Errors rates 8-15%
- Mostly indels
- 1 read = 1 full length transcript

Can SRCL be adapted to retrieve similarity between long reads representing complex transcripts?

→ Use small k value (15) and low percent of shared k-mers

Comparison to state of the art tools with 10k **mouse ONT (1D only) transcriptome** reads:

	Minimap	GraphMap	**SRCL**
Recall	55%	45.7%	**60.2%**
Precision	99.7%	**99.9%**	97.7%
Time(m:s)	00:09	00:05	2:51
Mem (G)	2.288	6.039	3.628

Future works: **post-processing for better precision, work on time consumption**

References

★Marchet et al, A resource-frugal probabilistic dictionary and applications in bioinformatics, 2017 (submitted to DAM)
★Meng et al, A transcriptomic approach to study marine plankton holobionts,2017 (International Conference on Holobionts)

Tool

https://github.com/GATB/short_read_connector

Acknowledgments

☺ JM Aury, C Da Silva and Aster ANR
☺ F Not, E Corre and S Le Crom
☺ BBHash Team
☺ Genouest Cluster

JOBIM Lille 3-6 juillet 2017 · Université de Rennes 1 · IRISA Inria · GenScale