Des lectures aux transcrits : méthodes *de novo* pour l'analyse du séquençage des transcriptomes de deuxième et troisième génération

Camille Marchet
sous la direction de Pierre Peterlongo

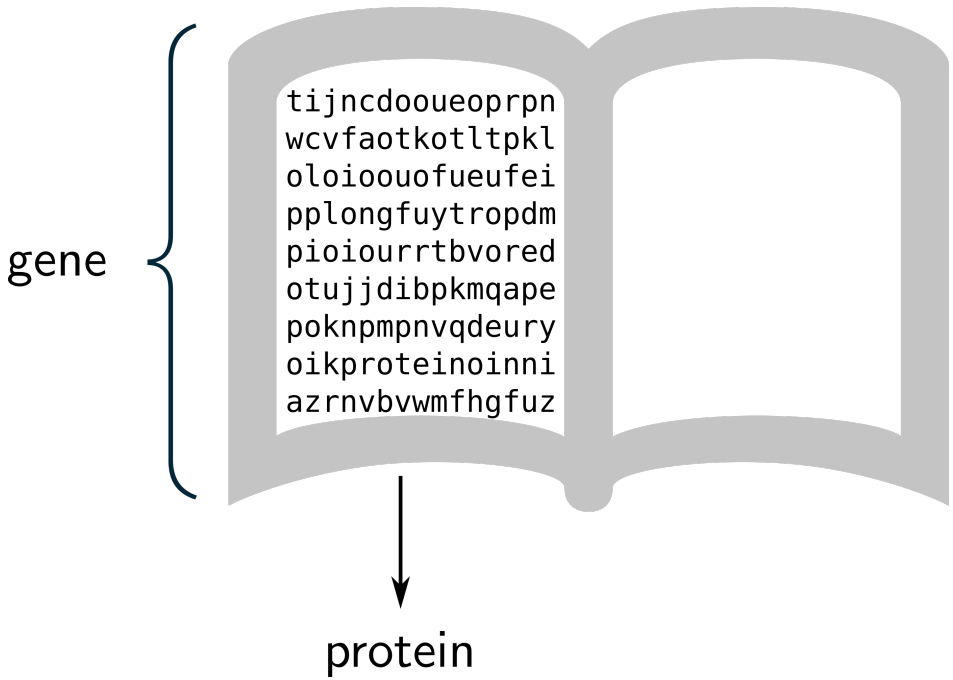
Congrès annuel de la SIF

5 février 2020

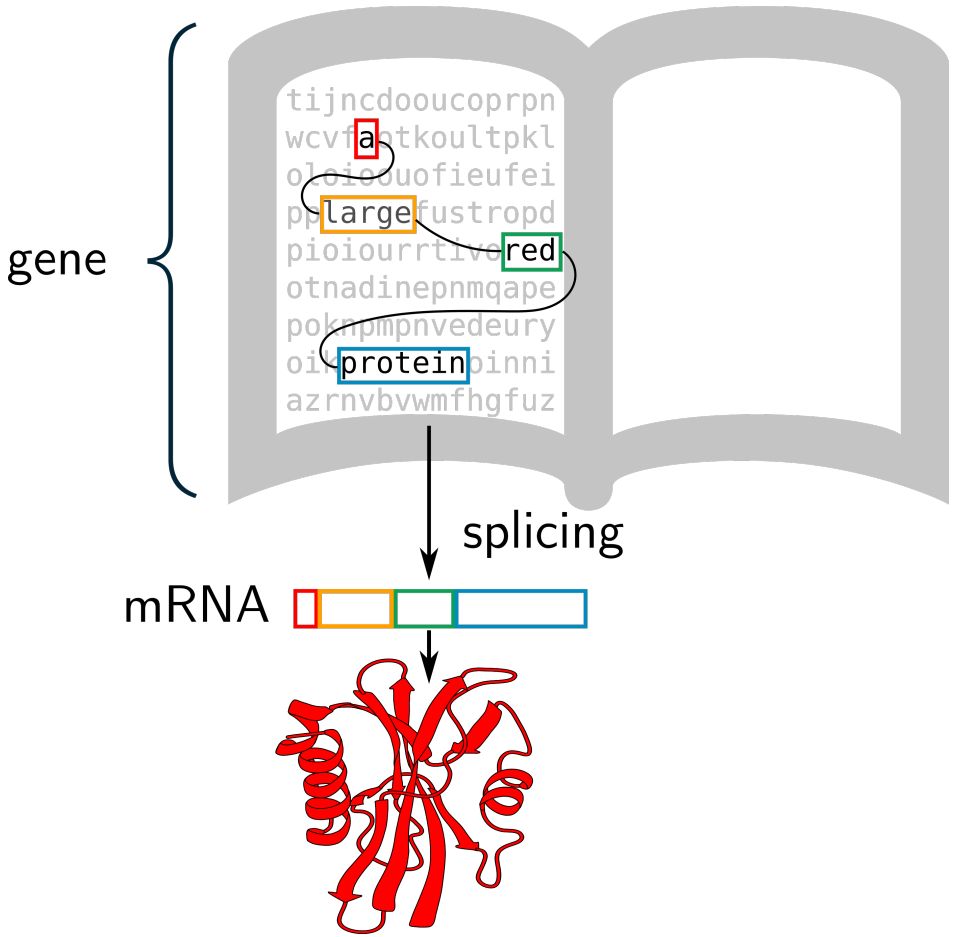
Introduction

- 2013 INSA de Lyon/Université Claude Bernard Lyon 1
  - Bioinformatics and Modeling
  - Ecology, Evolution, Biometrics
- 2013-15 Engineer in ERABLE team (Inria) - LBBE Lyon
  - Software development
- 2015-18 PhD thesis in Informatics, GenScale team (Inria) - IRISA Rennes, Université de Rennes
  - Algorithms for RNA sequences
- 2018- Postdoc in BONSAI team (CNRS) - CRIStAL Lille
  - Data structures for sequence bioinformatics

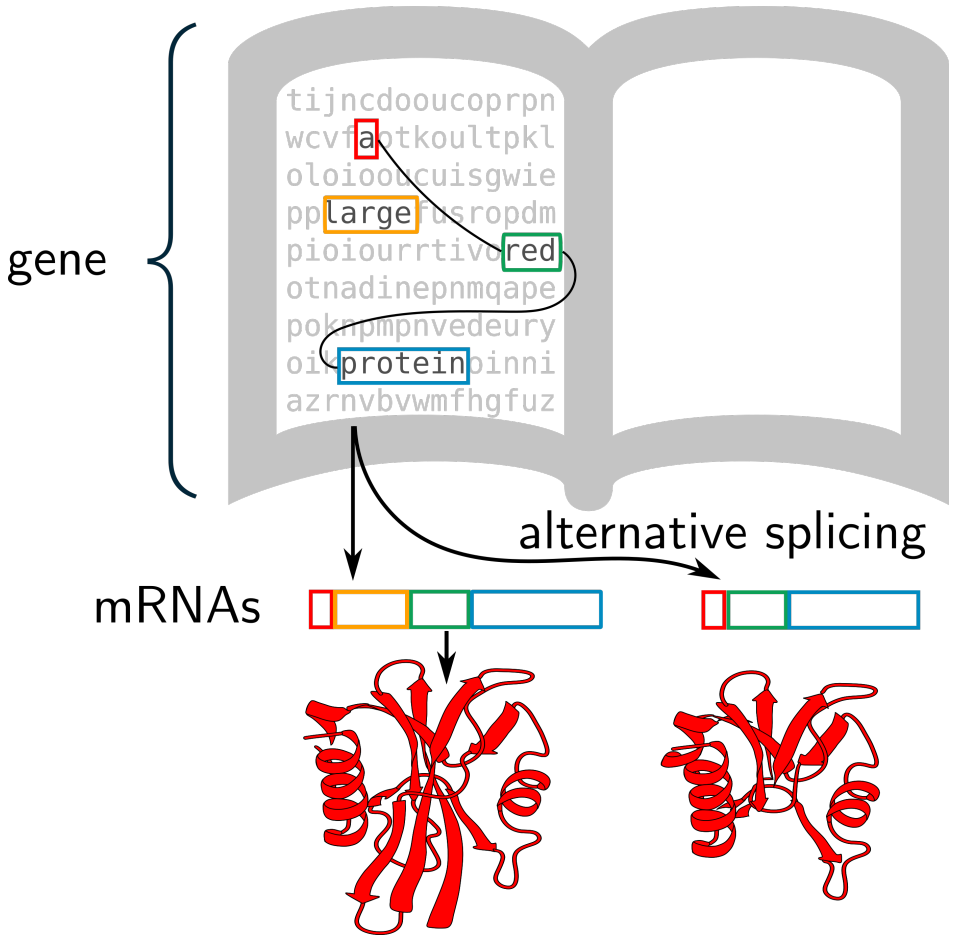
Introduction - Messenger RNAs



Introduction - Messenger RNAs

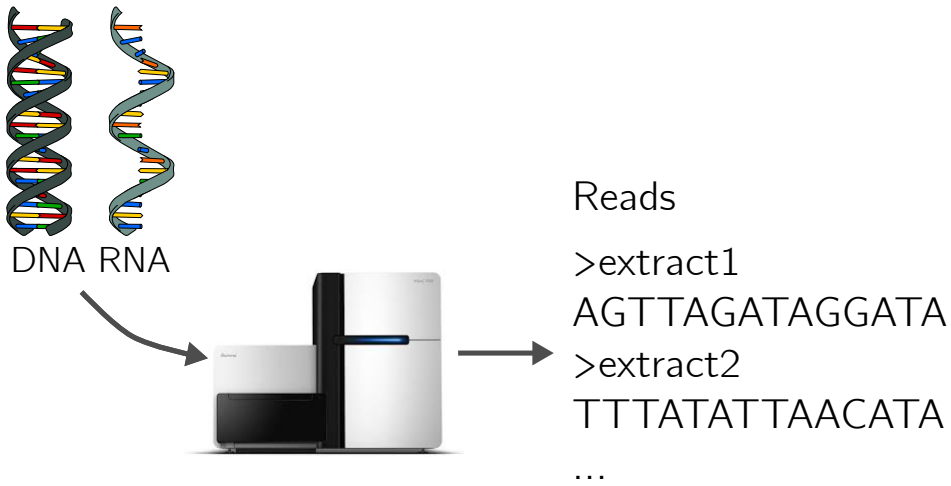


Introduction - Variability in messenger RNAs



There is a **combinatorial aspect** in messenger RNAs in eukaryotes (typically, mammals or plants)

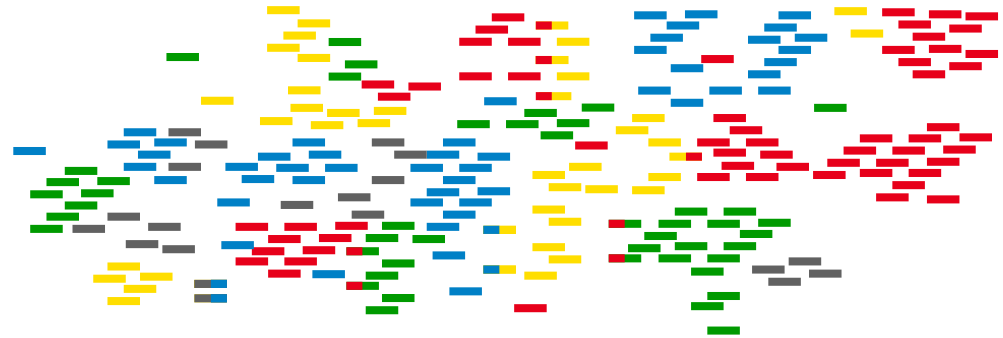
Introduction - Sequencing technologies

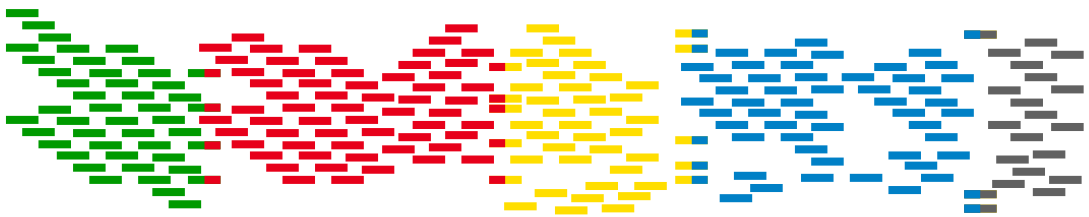


- **Reads** are substrings from DNA/RNA, in a 4 lettres alphabet (called **bases** or **nucleotides**)
- One dataset can contain billions of reads
- Today we can sequence >1 petabases a day

Introduction - Short reads

reads: shuffled short sequences (100 bases)

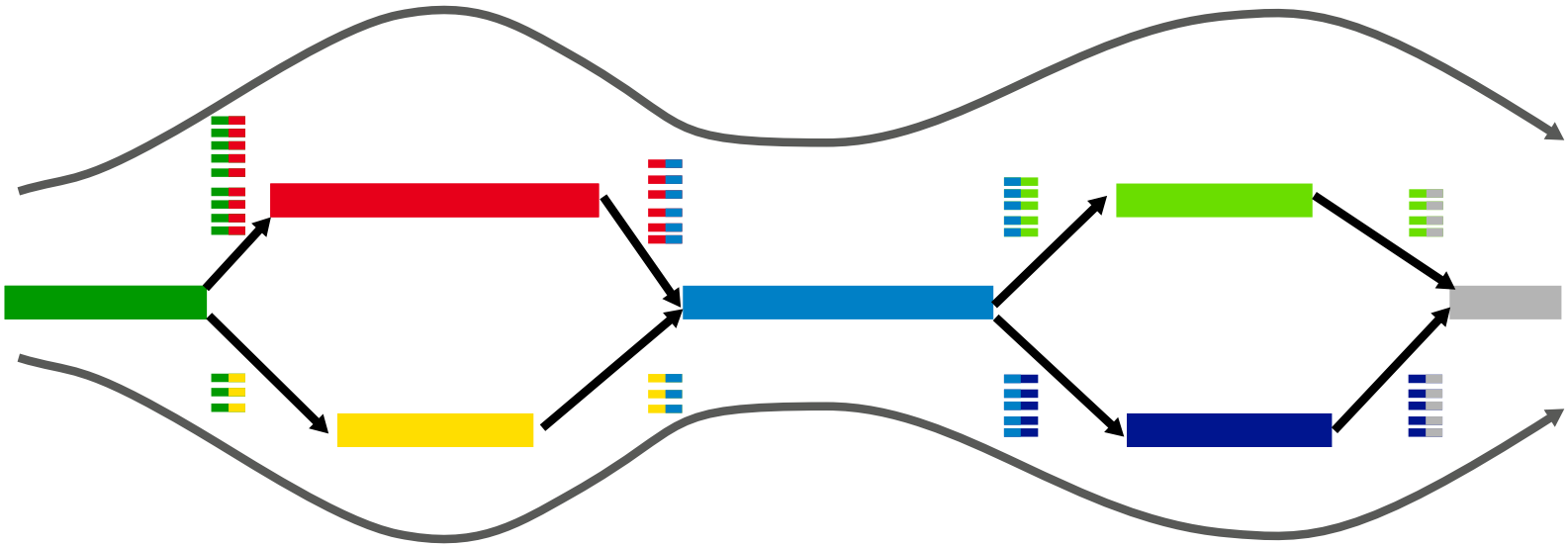




find order and overlaps (graph strategies)

final sequences (a few thousand bases)

Introduction - Issue with short reads

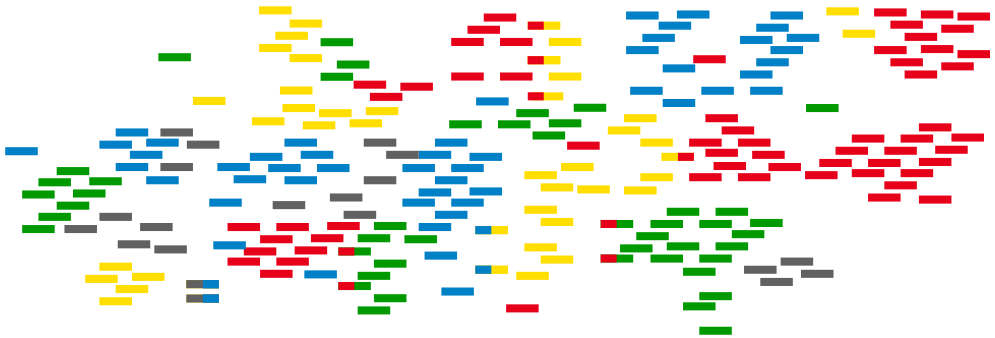


exist in experiment

output

Introduction - Long reads

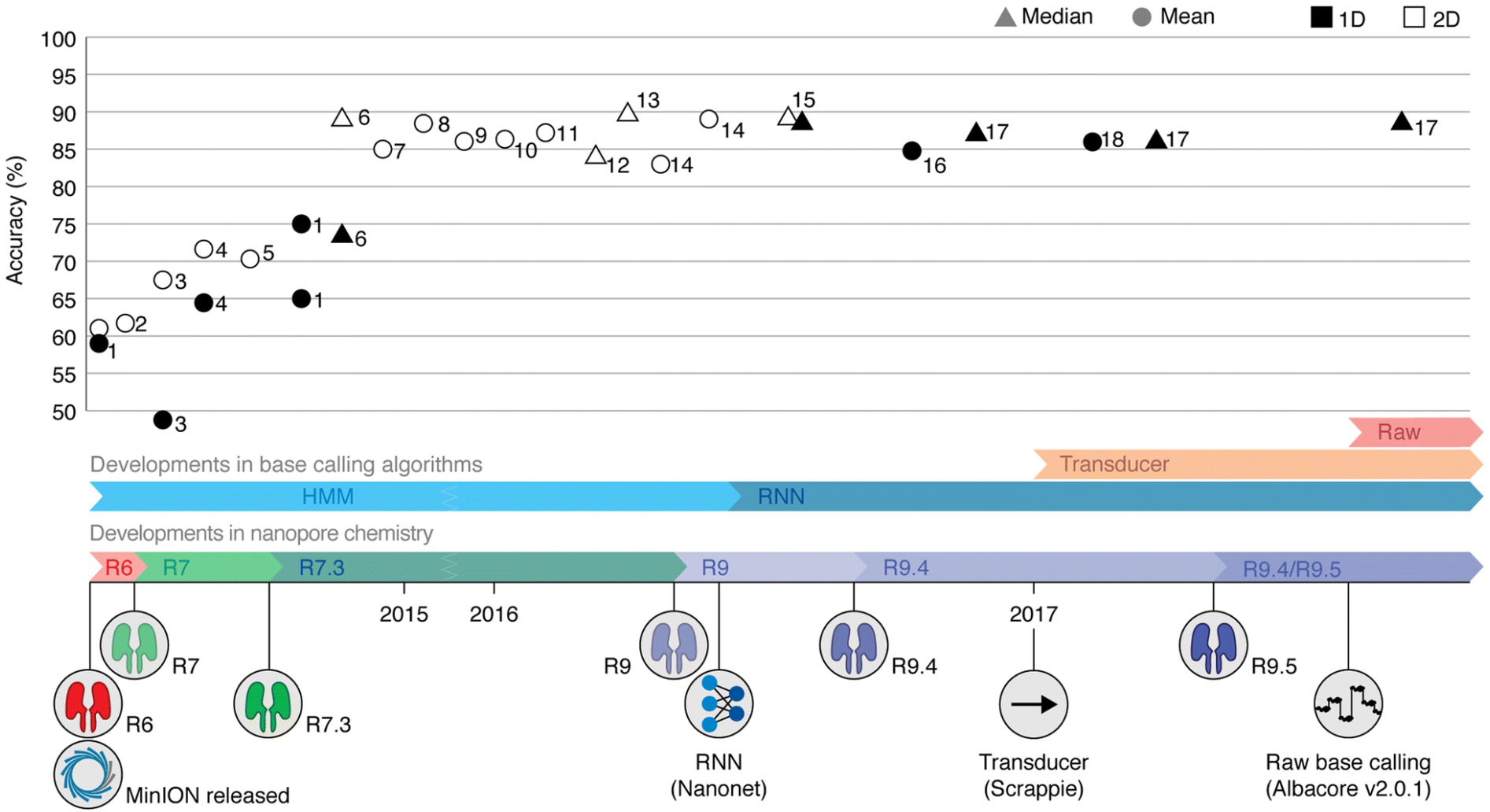
short reads



long reads

real sequence

Introduction - Issue with long reads

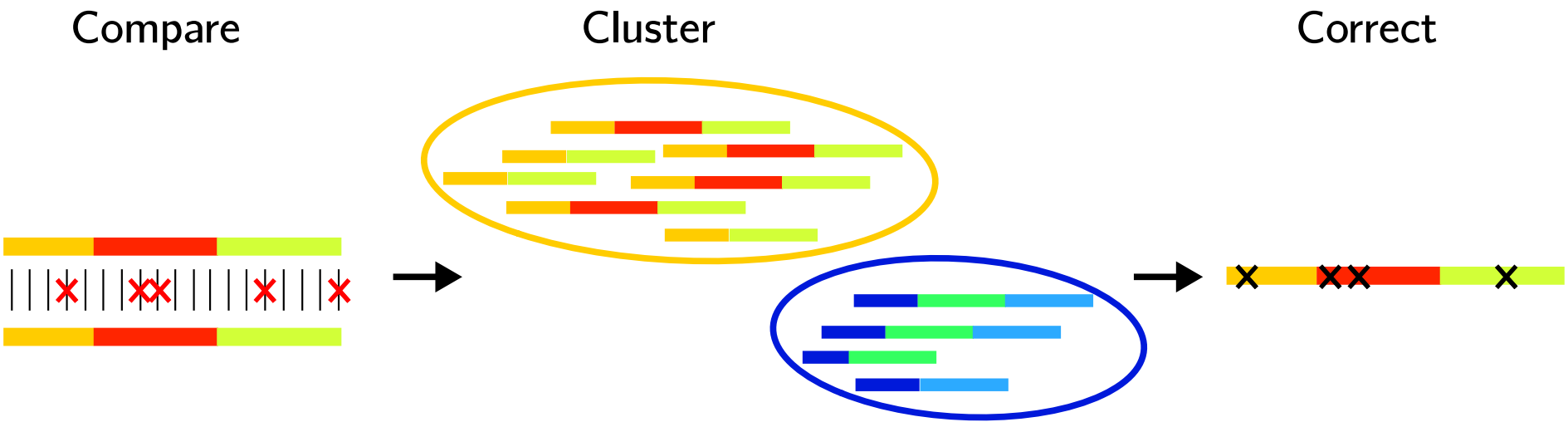


read AGGTGAATT▪GC
original sequence AG▪TGACTTTGC

Introduction - Five challenges I was interested in

- Gene expression (different levels of RNA molecules)
- Combination of mRNA for a given gene
- Errors in long reads
- Scalability (millions-billions of reads)
- *De novo* (do not rely on sequences that are already known)

Thesis outline



Sequence comparison - Compare RNA strings

Dynamic programming algorithm (Needleman & Wunsch, Smith & Waterman)

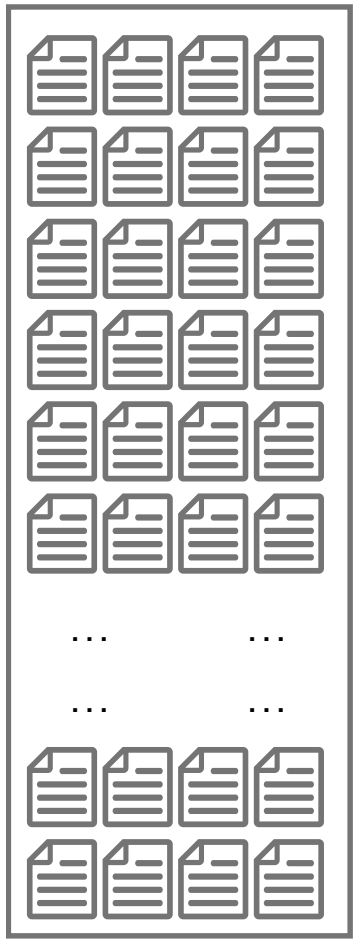
M

		-	A	A	C	C
	-	0	-1	-2	-3	-4
	A	-1	0	-1	-2	-3
L	A	-2	-1	0	-1	-2
	G	-3	-2	-1		
	C	-4	-3	-2		

O(L x M)

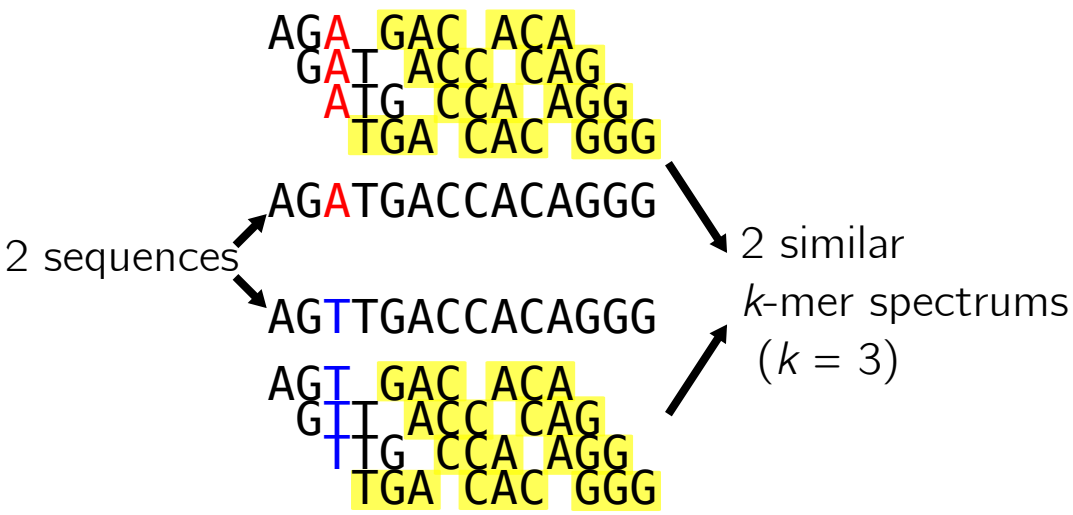
```
AACC
|| |
AAGC
```

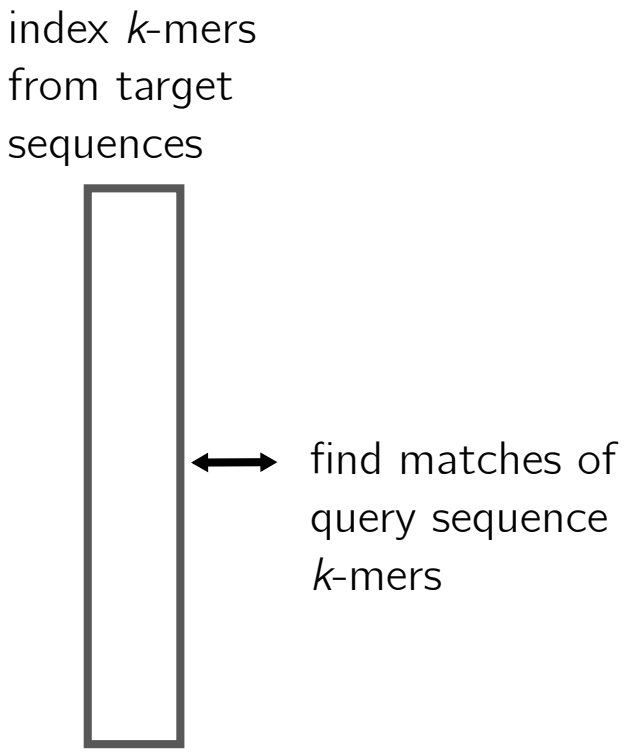
non scalable to this problem:



Sequence comparison - Heuristics to compare DNA/RNA strings

Use *k*-**mers**: words of size *k* in the sequences

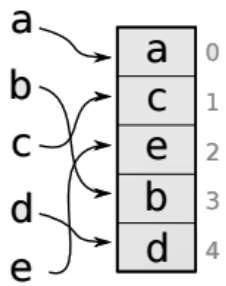




Sequence comparison - Our solution to compare sequences at scale

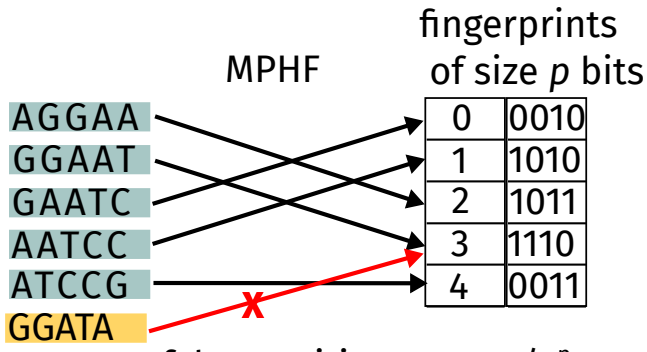
1- Minimal perfect hashing

(no collisions, |image|=|input|)



[Limasset et al. 2017]

2- QUASI-DICTIONARY

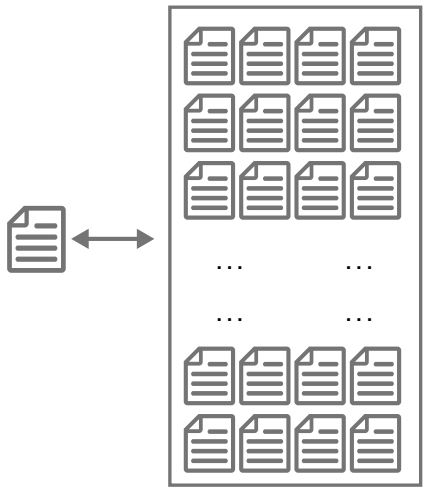


false positive rate = $1/2^p$

Memory consumption

MPHF: half a byte
default p = 12
~ 2 bytes/k-mer for a 0.02% FP rate

3- SHORT READS CONNECTOR

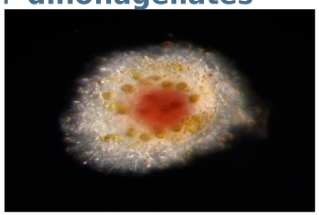


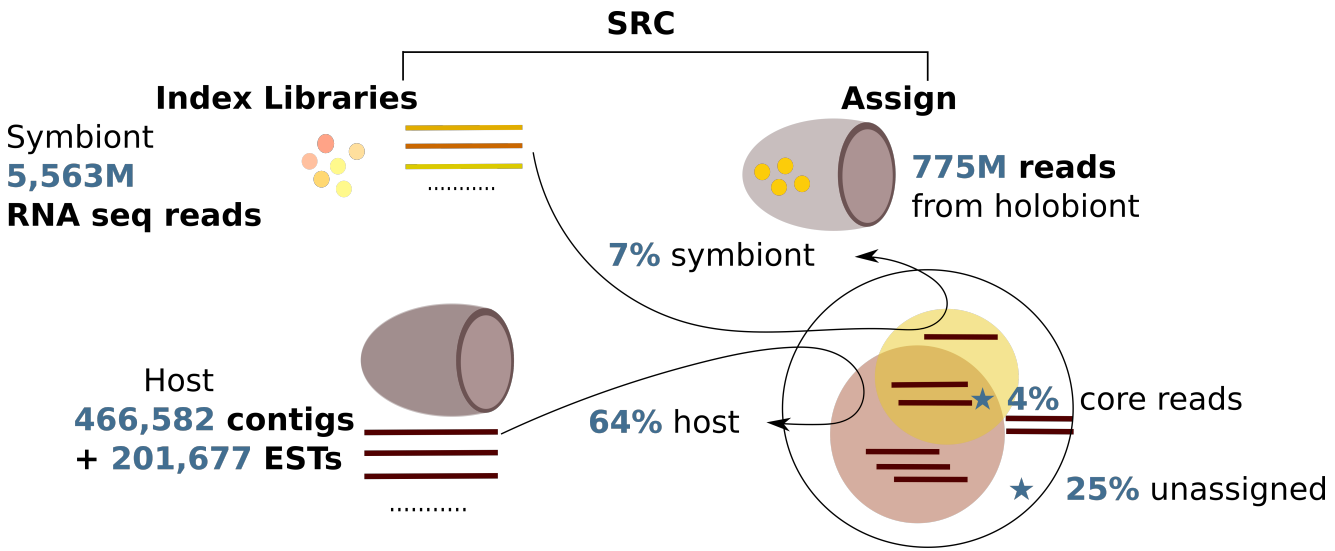
[Marchet et al. 2018]

Sequence comparison - Application to plankton

Model 3 (M3)

Radiolaria - Dinoflagellates:
Eucyrtidium acuminatum
+ dinoflagellates





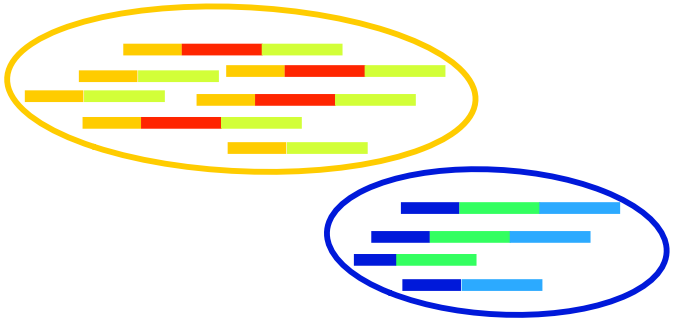
$\sim$ 7 hours and 40GB RAM [Meng et al. 2018]

Large scale sequence comparison

- **Fast-expanding field**: 18 papers and counting since SRC (data-structure improvements) for the indexation of collection of datasets problem
- We proposed an **exact data-structure** on top of the quasi-dictionary work
- Works well for short reads, still an ongoing work for long reads

Sequence clustering - The case of RNA long reads

GOAL: 1 Cluster per gene

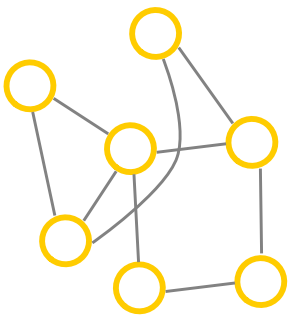


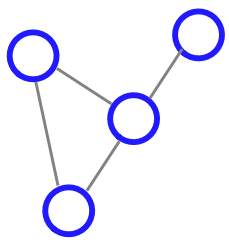
DNA: long sequences ($> 10^3$ bases)
similarity **reported** by current methods

RNA: shorter shared subtrings ($< 10^2$ bases)
not always reported by current methods

Clustering input:

nodes = reads
edges = reported similarity



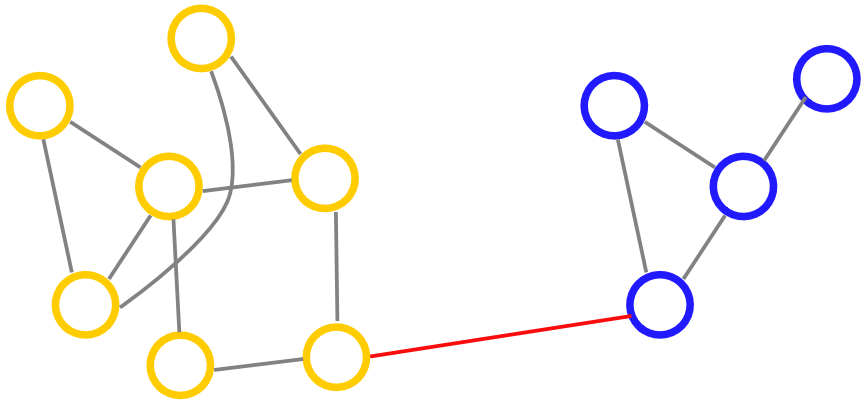


Sequence clustering - Community finding algorithms

Clustering input:

nodes = reads

edges = reported similarity



Expected: missing edges (badly connected quasi-cliques)
very heterogeneous cluster sizes (gene expression)
some spurious edges

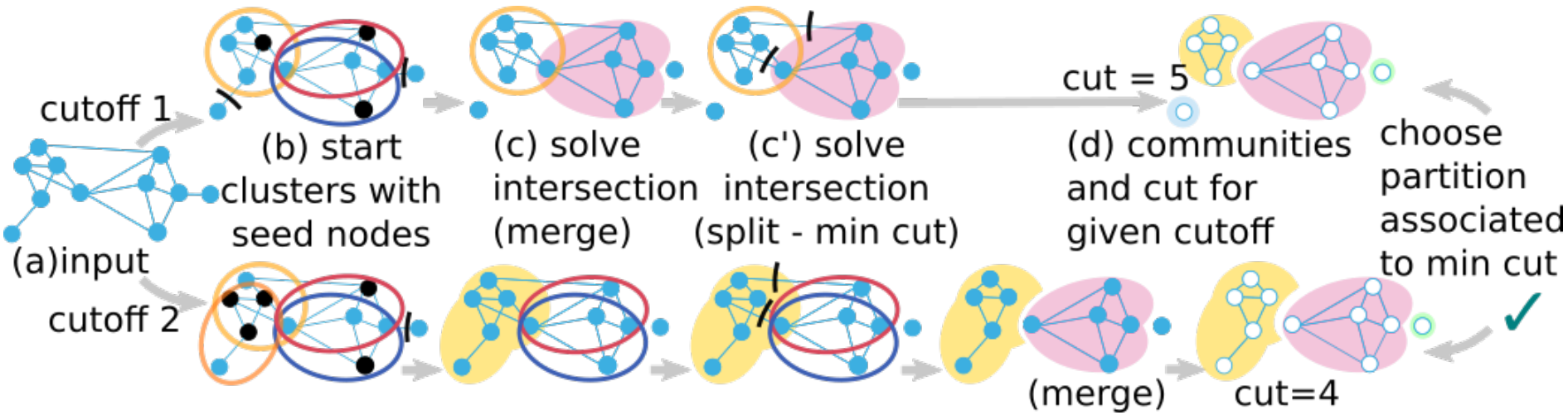
	Recall (%)	Precision (%)	F-measure (%)	Jaccard index
Connected component	75.74	5.614	13.62	$7.3E^{-4}$
Modularity	60.70	71.16	65.51	$9.7E^{-2}$
CPM5	79.00	69.35	73.86	$3.5E^{-1}$
CPM50	49.21	**89.92**	63.60	$7.6E^{-2}$
Louvain	**88.58**	14.91	25.53	$1.1E^{-3}$

Sequence clustering - Our solution

Intuition:

- ideal case = a clique per gene. Use the **clustering coefficient** θ as a connectivity metric
- we don't know in advance the number of clusters (k)
- mostly biologically sound edges find a **minimum** k**-cut**, NP-hard for ≥ 3 [Dahlhaus et al. 1994]
- approximation of the solution: explore a restricted space for k
- explore **local cutoffs** for θ

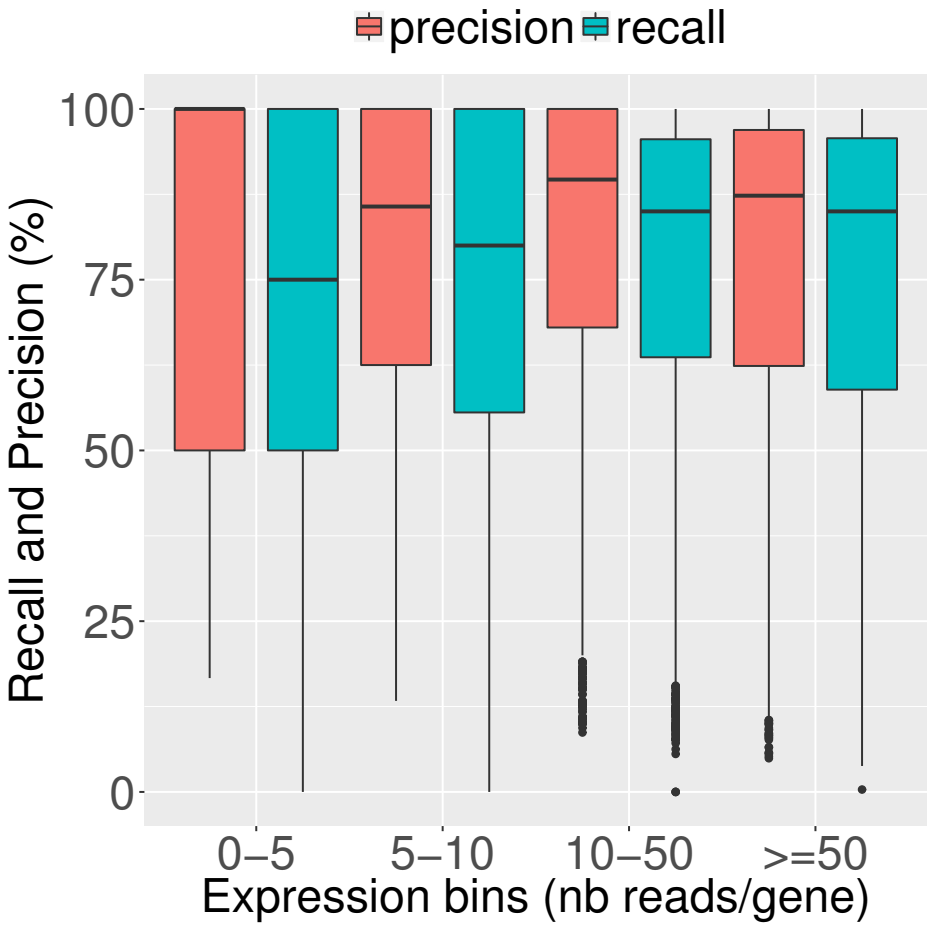
Sequence clustering - Our solution

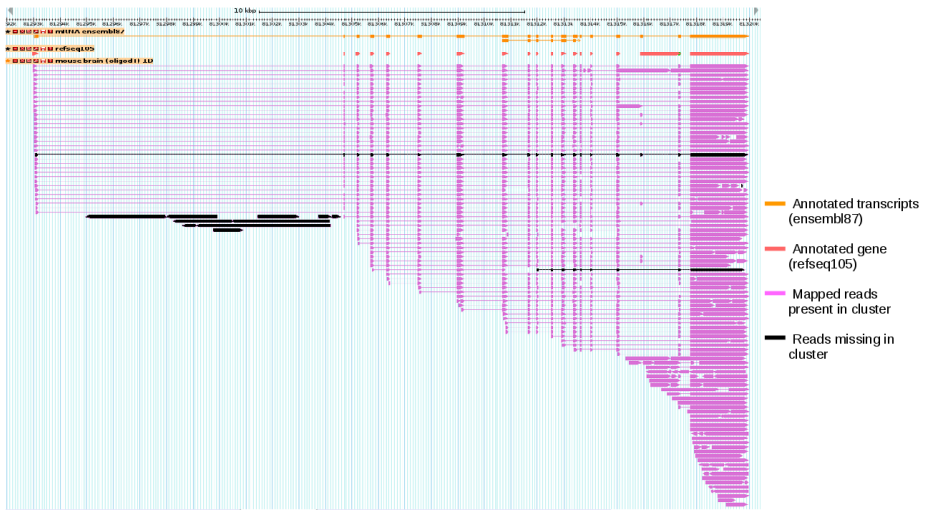


Software CARNAC-LR [Marchet et al. 2018]

Sequence clustering - Application to mouse data

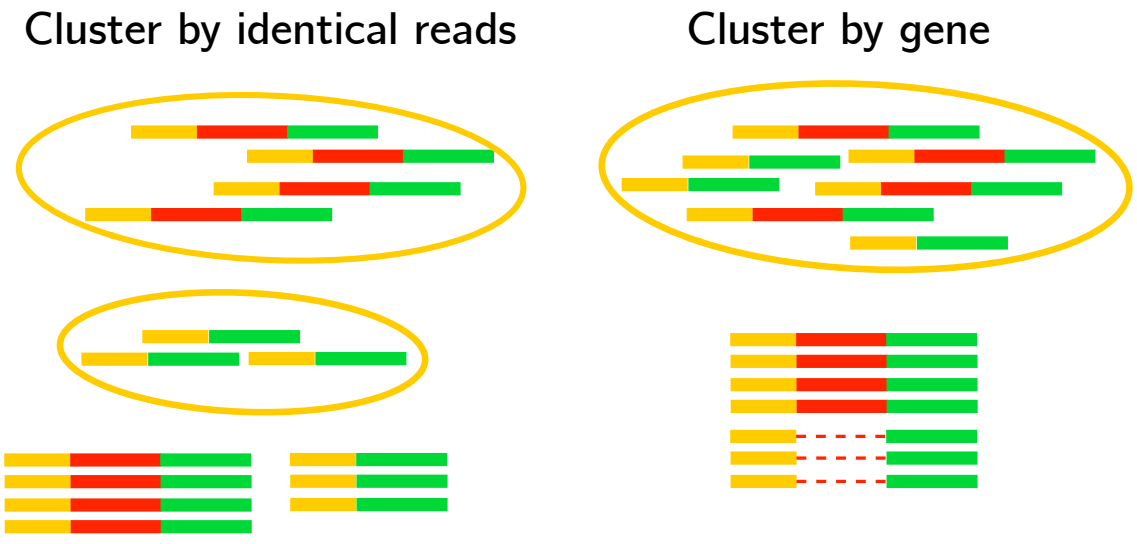
$\sim$ 500,000 long reads from a mouse



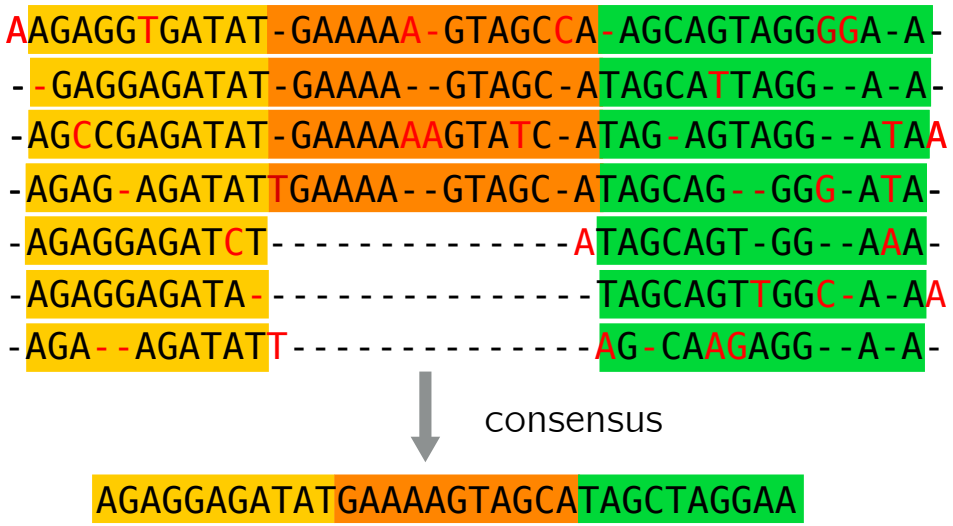


RNA Long Read Sequence clustering

- Independently, another similar method emerged just after we published [Sahlin et al. 2019]
- Next step: scalability, correction

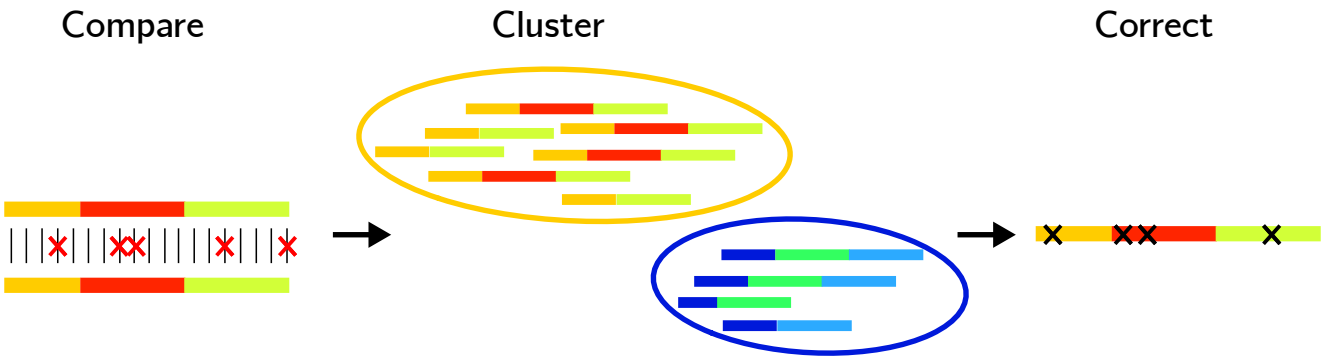


Correction - Segmented multiple sequence alignment



- We used the segmented multiple sequence alignment for long reads in two correction-related articles [Marchet et al. 2020, Morisse et al. 2019], notably we **corrected human reads**
- Very recently: a preprint with the same idea (gene clustering + correction) [Sahlin et al. 2020]
- Long read correctors are not well-tailored for RNA [Lima et al. 2019]

Conclusion



- Set of sequences indexation
- Sequence clustering
- Sequence correction by multiple alignment

Acknowledgments

SIF members / GenScale + Dyliss + Genouest teams / ANR ASTER / ANR Hydrogen / IBPS / Genoscope / TIBS team