Optimizing early steps of long-read genome assembly

Pierre MARIJON, Maël KERBIRIOU, Jean-Stéphane VARRÉ, Rayan CHIKHI

November 20, 2018

盆栽 team, Lille

What's a long-read?

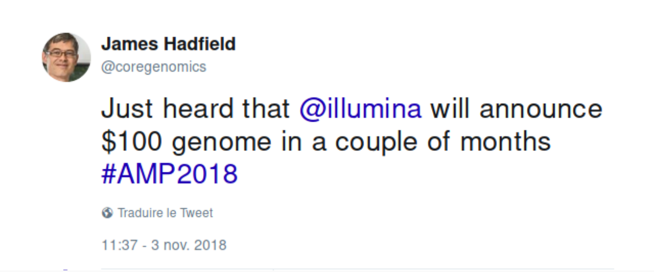
Third generation reads are :

- Long > 10kb [1]
- Erroneous $\approx 16\%$ [1]
- Chimeric [2]

[1] Jain et al. 2018

[2] Laver et al. 2016

Sequencing faster, cheaper, stronger

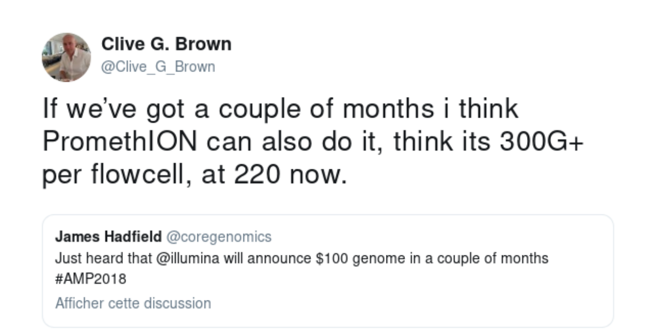


James Hadfield
@coregenomics

Just heard that @illumina will announce $100 genome in a couple of months #AMP2018

Traduire le Tweet

11:37 - 3 nov. 2018



Clive G. Brown
@Clive_G_Brown

If we've got a couple of months i think PromethION can also do it, think its 300G+ per flowcell, at 220 now.

James Hadfield @coregenomics
Just heard that @illumina will announce $100 genome in a couple of months #AMP2018
Afficher cette discussion

What we can do with long-read?

By mapping against reference:

- read correction
- variant calling
- . . .

against themselves:

- self correction
- assembly
- . . .

Long-read mapping

Many tools :

- minimap[2]
- mhap
- ngmlr
- graphmap
- daligner
- . . .

Some output format:

- MHAP:

```
read1  read2  0.14  1955  0  998  20480  21581  0  45  19527  19801
```

- Pairwise Alignement Format:

```
read1  21581  998  20480  +  read2  19801  45  19527  1955  19482  255
```

- SAM

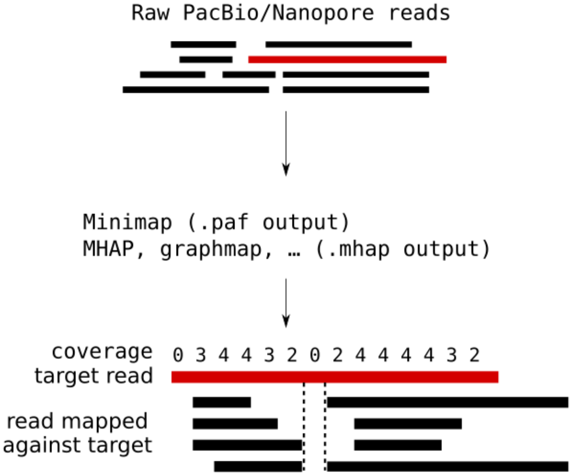
Correction?

Correction involves a lot of operations and costs time and memory.

I just want to detect chimeras.

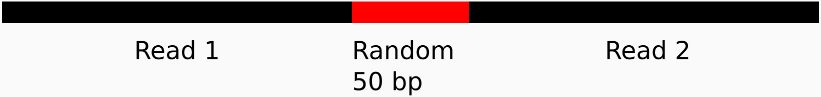
Yet Another Chimeric Read Detector

Chimeric read: when a part of the read is not well supported (i.e. covered) by other reads of the dataset.



Yet Another Chimeric Read Detector

Test dataset: 20x synthetic long read[3] of *T. roseus*



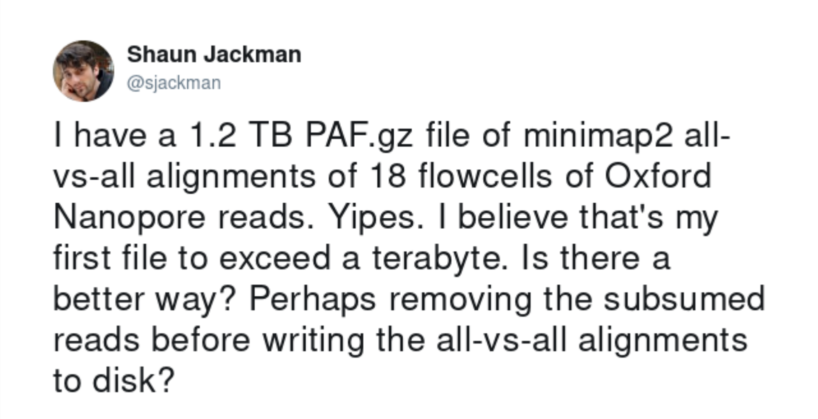
[3] LongISLND with pacbio error model

Yet Another Chimeric Read Detector

	minimap2 + yacrd	DAScrubber[4]
wallclock time (seconds)	48.13	365.79
precision	100.00%	87.70%
sensitivity	70.34%	71.16%

[4] run by https://github.com/rrwick/DASCRUBBER-wrapper

Another trouble: the disk space



Shaun Jackman
@sjackman

I have a 1.2 TB PAF.gz file of minimap2 all-vs-all alignments of 18 flowcells of Oxford Nanopore reads. Yipes. I believe that's my first file to exceed a terabyte. Is there a better way? Perhaps removing the subsumed reads before writing the all-vs-all alignments to disk?

18 flowcells produce ≈ 180Gb-540Gb

A summary of troubles and some possible solutions:

`https://blog.pierre.marijon.fr/binary-mapping-format/`

Filter Pairwise Alignment

FPA can filter on:

- type :
    - containment
    - internal match
    - dovetails
- self match
- overlap length
- read match against a regex

FPA can rename your read, compress (gzip, bzip, lzma) and convert your pairwise alignment in an overlap graph (GFA1)

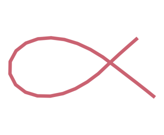
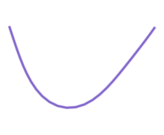
Filter Pairwise Alignment

	wallclock time (s)	output length (Mb) / % space saved	throughput (kb/s)
minimap2	866	565	652.320
minimap2 + fpa no filter	869	565 (0%)	650.047
minimap2 + fpa ovl length > 2000	868	452 (20%)	520.468
minimap2 + fpa dovetails only	869	401 (29%)	462.007

Dataset: SQK-MAP-006 2D nanopore read

`http://lab.loman.net/2015/09/24/first-sqk-map-006-experiment/`

Filter Pairwise Alignment

	minimap2	minimap2	
	+ miniasm	fpa + miniasm	diff
PAF file size (Mb)	565	452	-20%
assembly time (s)	6.5	6	0.5
assembly result			∅

Dataset: SQK-MAP-006 2D nanopore read

`http://lab.loman.net/2015/09/24/first-sqk-map-006-experiment/`

Conclusion

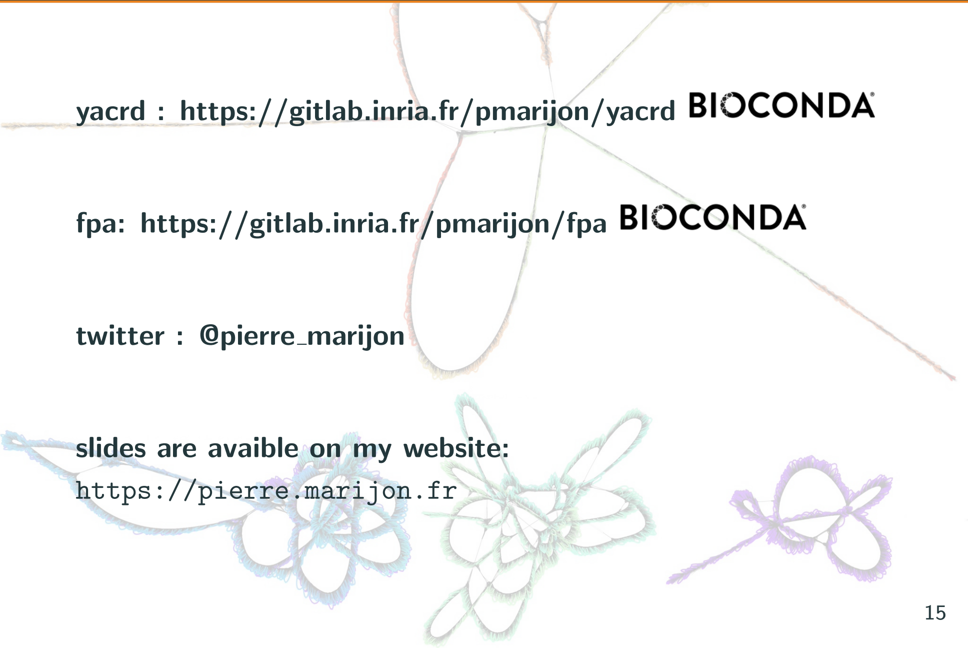
What we have:

- more and more third generation sequencing data
- analyses generate even more intermediate data
- with simple algorithms we can save time and space

What we need:

- compressed pairwise alignement format
- to detect more precisely poor quality regions

Questions?



yacrd : https://gitlab.inria.fr/pmarijon/yacrd BIOCONDA

fpa: https://gitlab.inria.fr/pmarijon/fpa BIOCONDA

twitter : @pierre_marijon

slides are avaible on my website:
`https://pierre.marijon.fr`