Genome assembly with long reads

Christophe Klopp

http://bioinfo.genotoul.fr/
http://www.sigenae.org/

Outline

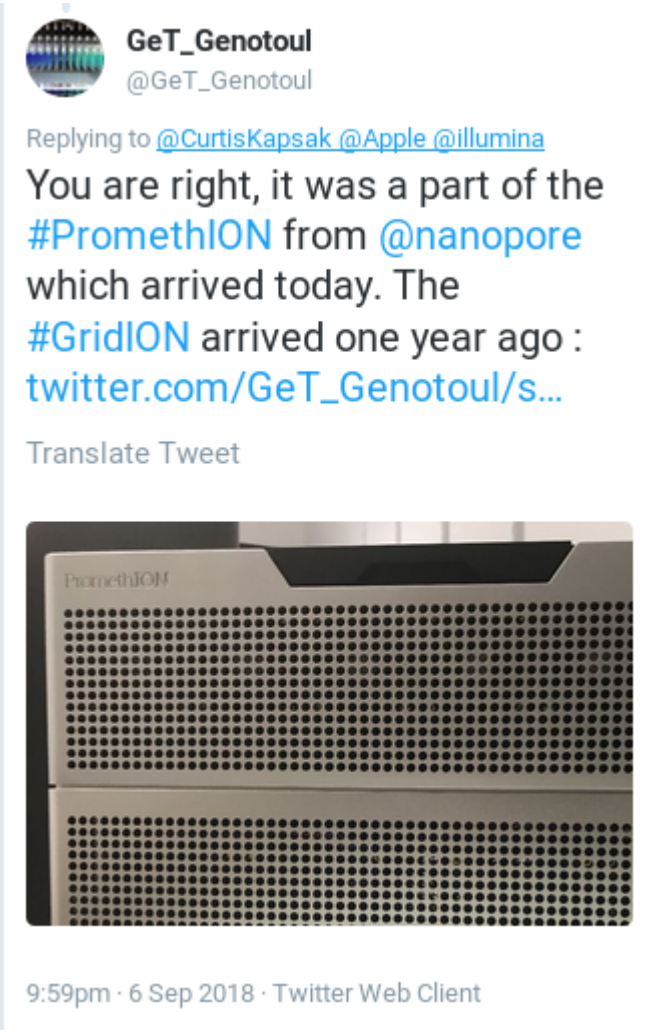
- Some news
- Nanopore read production and error patterns
- PacBio read production and error patterns
- Assembly strategies & results
- Assembly polishing
- Scaffolding
- Difficult to assemble genomes
- Conclusions

Oxford Nanopore PromethION

http://get.genotoul.fr/

Up to **150Gb** per flowcell.

48 flowcells will be able to run different samples in parallel.



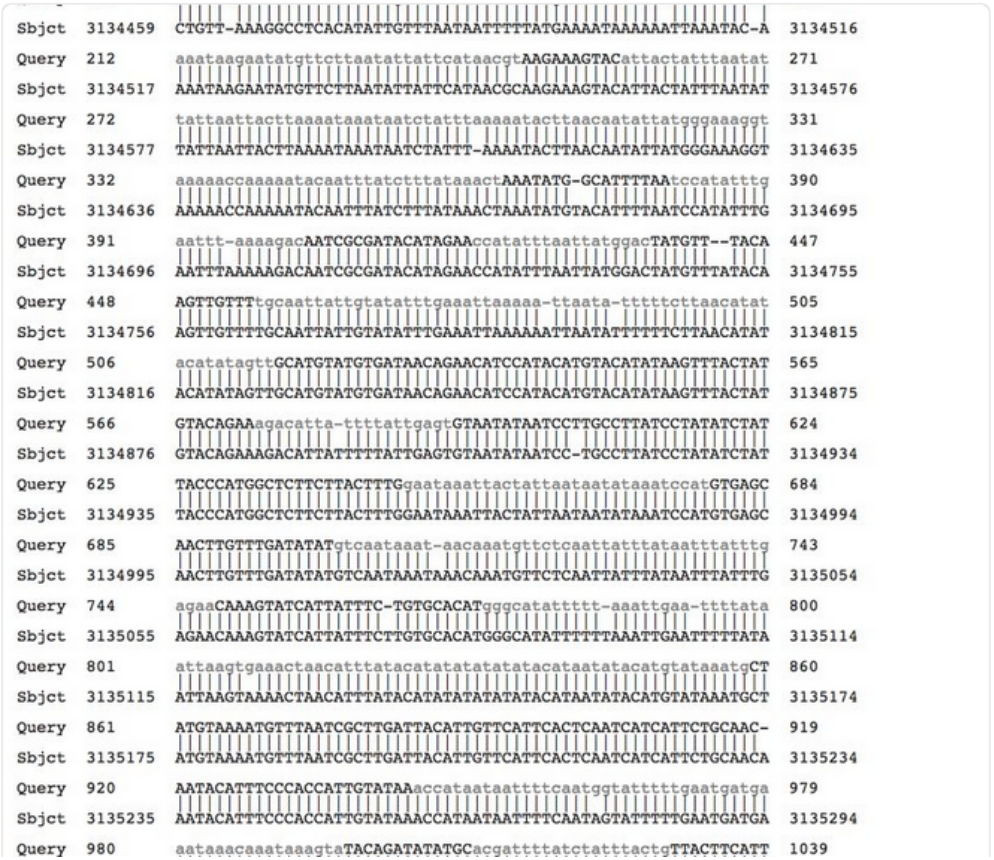
97 % accuracy for 1D^2

Oxford Nanopore a retweeté

Ola Wallerman @OWallerman · 2 juin

Alba v1.2 with 1d^2 calling now available, just in time. :) Best reads so far ~97% accuracy

À l'origine en anglais



PacBio Sequel

Sequel System: Superior accuracy, long reads, uniform coverage

The Sequel System is based on our proven Single Molecule, Real-Time (SMRT) Sequencing technology achieves:

- 10 Gb with 20 kb average read lengths for whole genome sequencing projects
- 20 Gb with 40 kb average read lengths for amplicon and RNA sequencing projects
- >99.999% (QV50) consensus accuracy with data free of systematic errors
- With an efficient <1 day workflow

REQUEST PRICING DOWNLOAD BROCHURE

https://www.pacb.com/products-and-services/sequel-system/

PacBio

nature genetics

Home | Current issue | Comment | Research | Archive ▾ | Authors & referees ▾ | About the journal ▾

home ▸ archive ▸ issue ▸ news and views ▸ full text

NATURE GENETICS | NEWS AND VIEWS

日本語要約

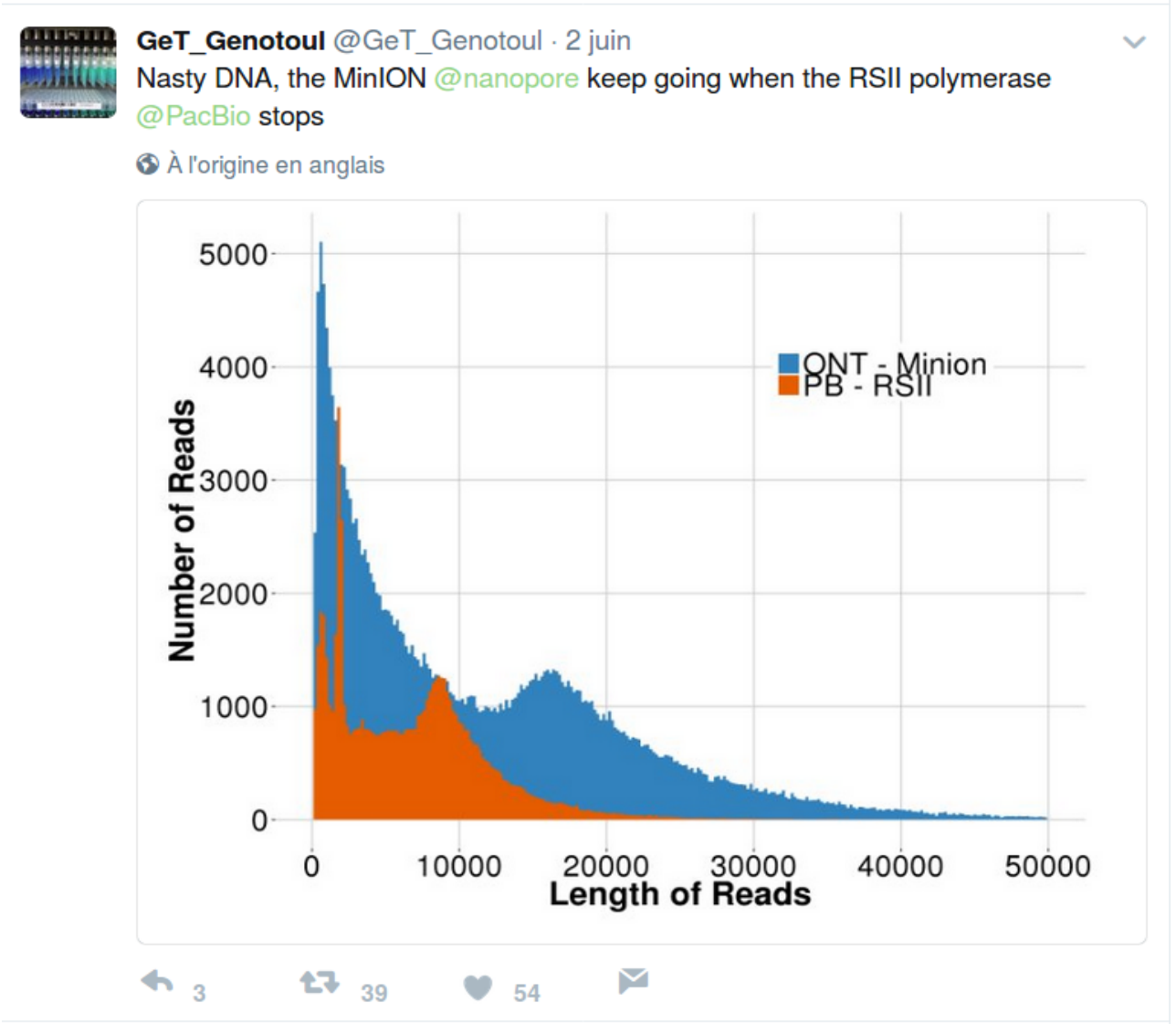
A golden goat genome

Kim C Worley

Nature Genetics **49**, 485–486 (2017) | doi:10.1038/ng.3824
Published online 30 March 2017

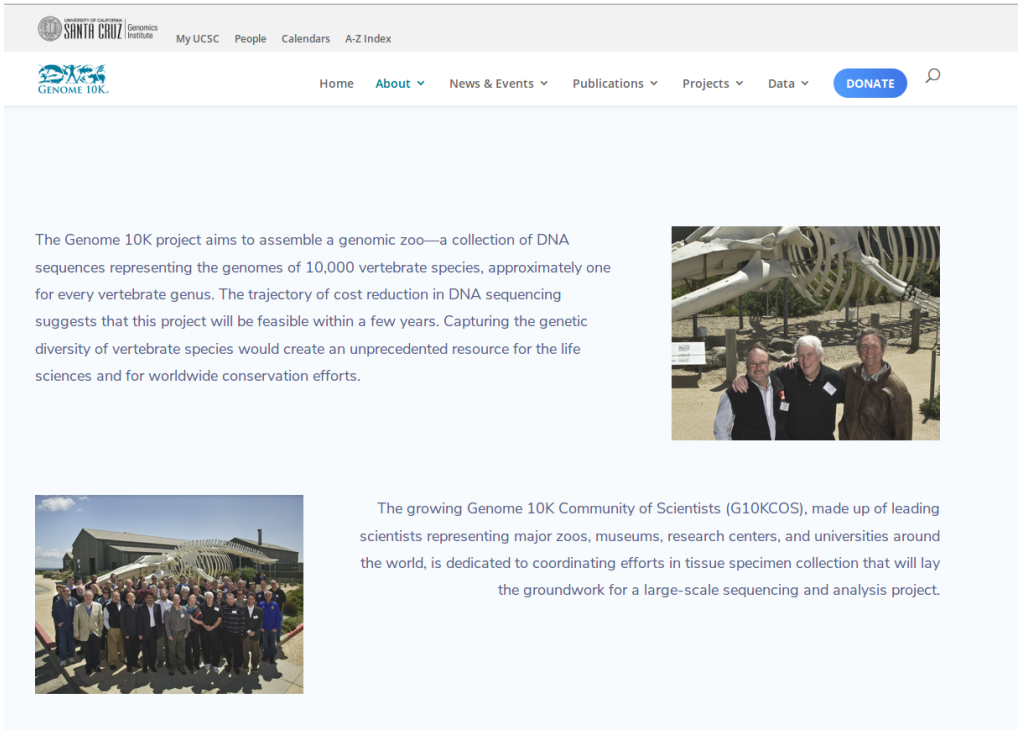
The newly described *de novo* goat genome sequence is the most contiguous diploid vertebrate assembly generated thus far using whole-genome assembly and scaffolding methods. The contiguity of this assembly is approaching that of the finished human and mouse genomes and suggests an affordable roadmap to high-quality references for thousands of species.

Same sample / RSII vs MinION

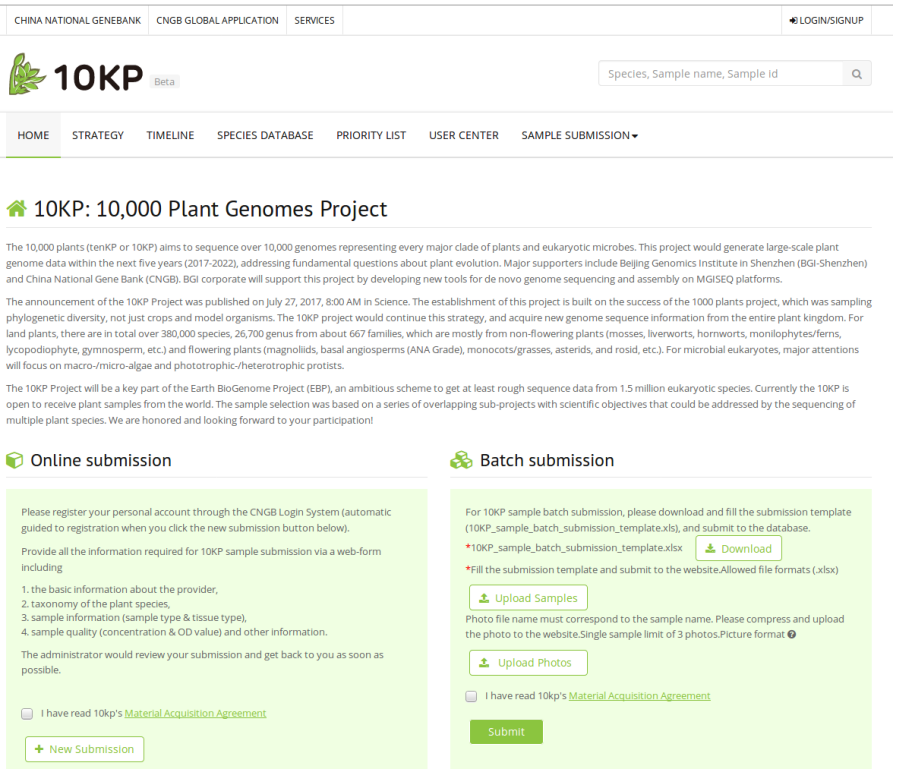


10K genome projects

https://genome10k.soe.ucsc.edu/



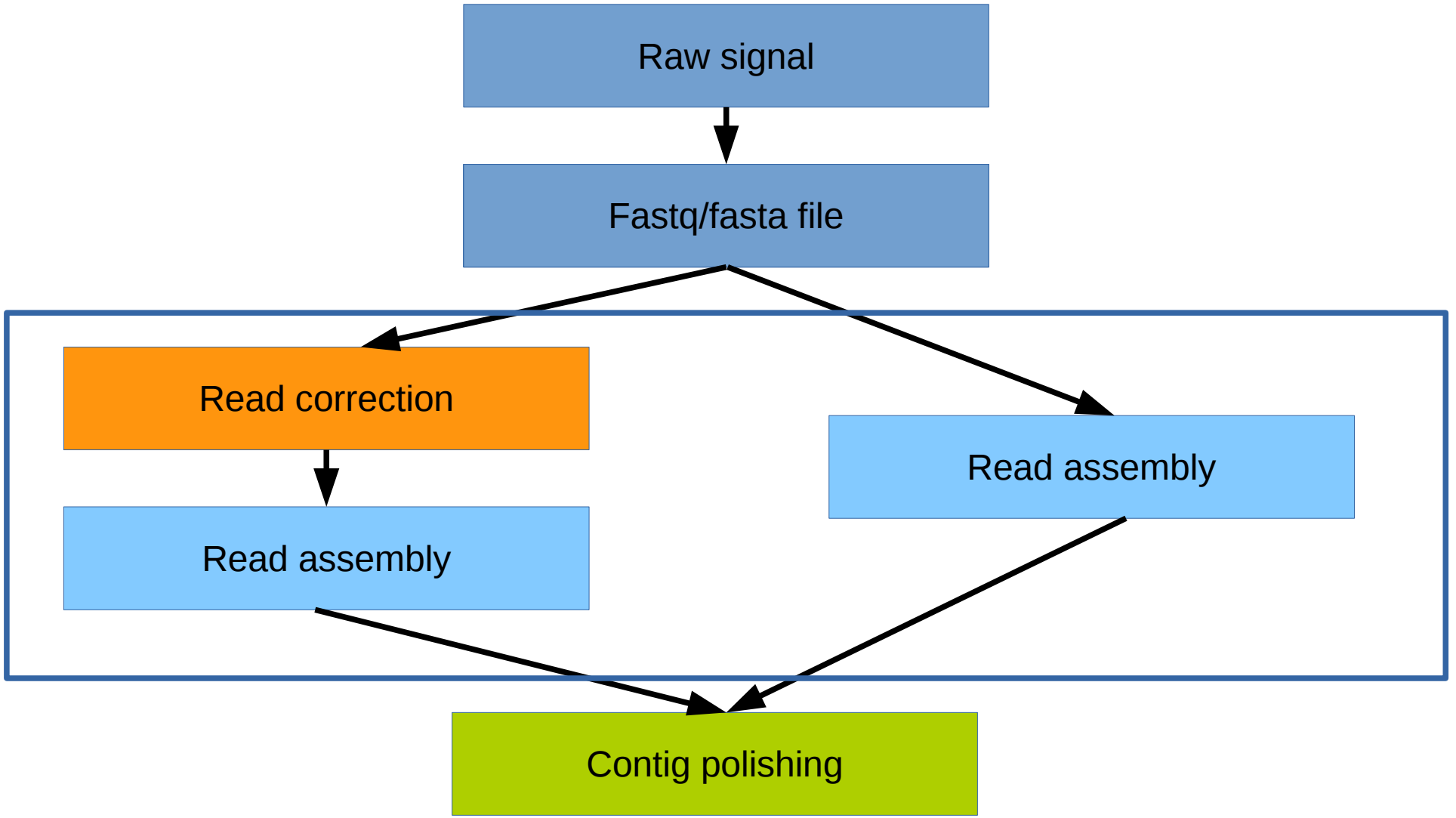
https://db.cngb.org/10kp/



Why use long reads to assemble genomes?

- Because
  - they bridge repetitions and build less fragmented genomes.
  - they come from technologies which do not amplify the DNA fragments and therefore have less coverage bias.
  - they are affordable.
  - they provide methylation information as well.

Assembly processing pipeline



Two technologies

Oxford Nanopore

MinION

GridION

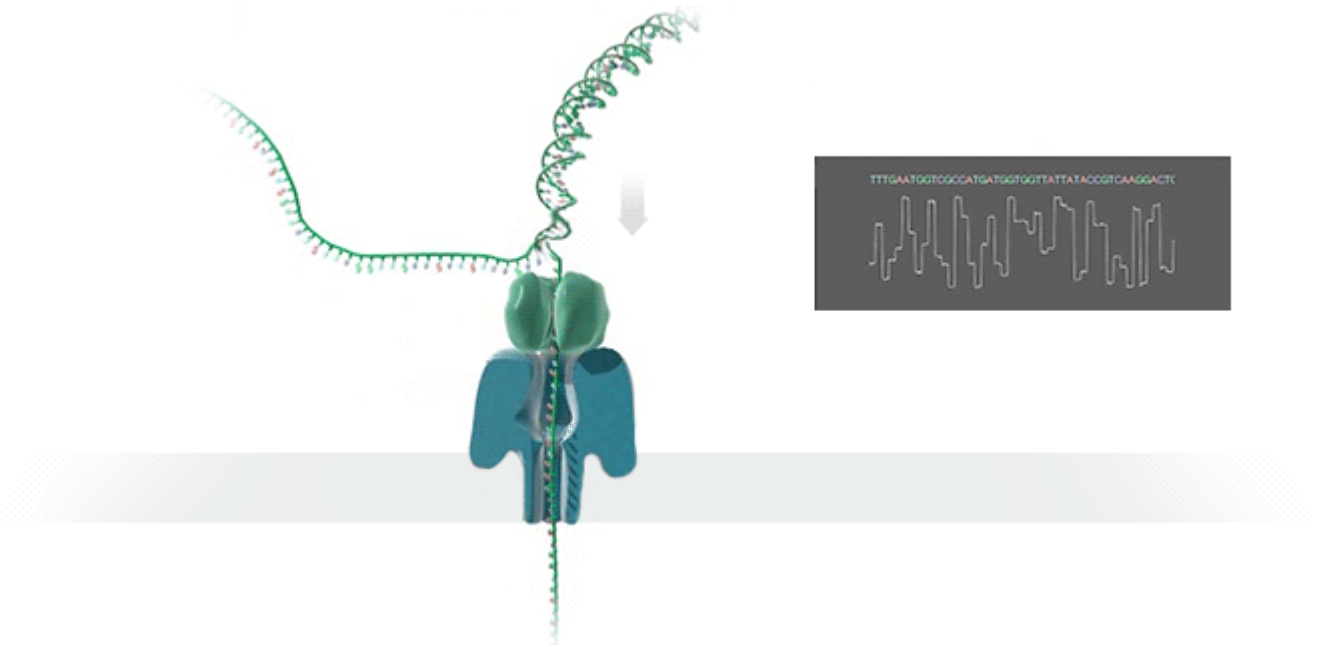
PromethION

Pacific BioScience

RSII

Sequel

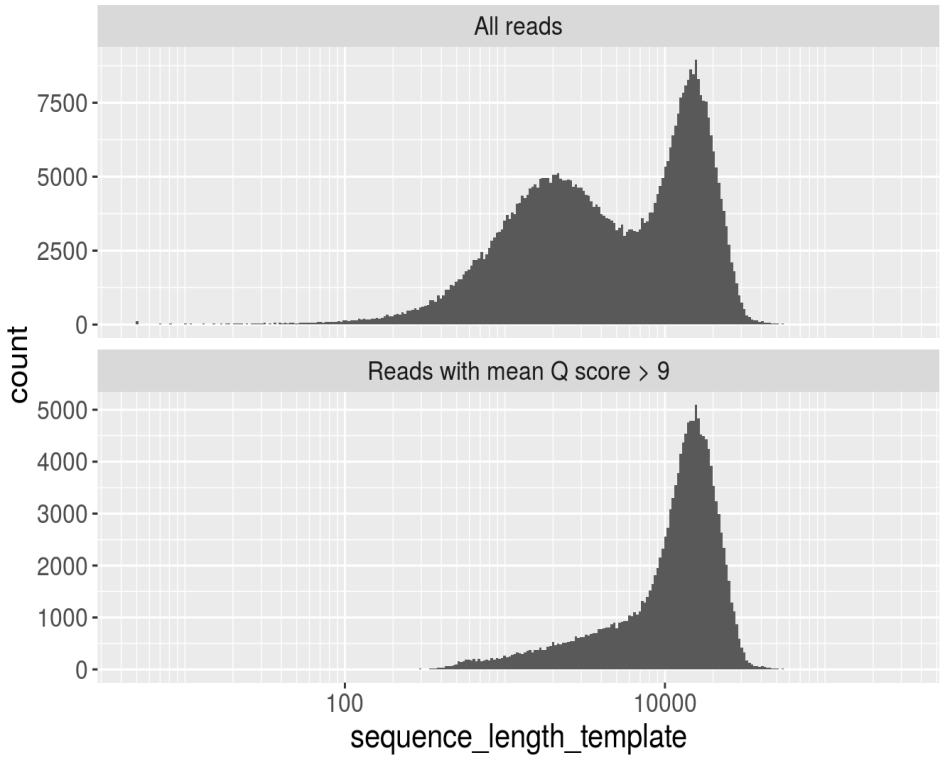
ONT technology



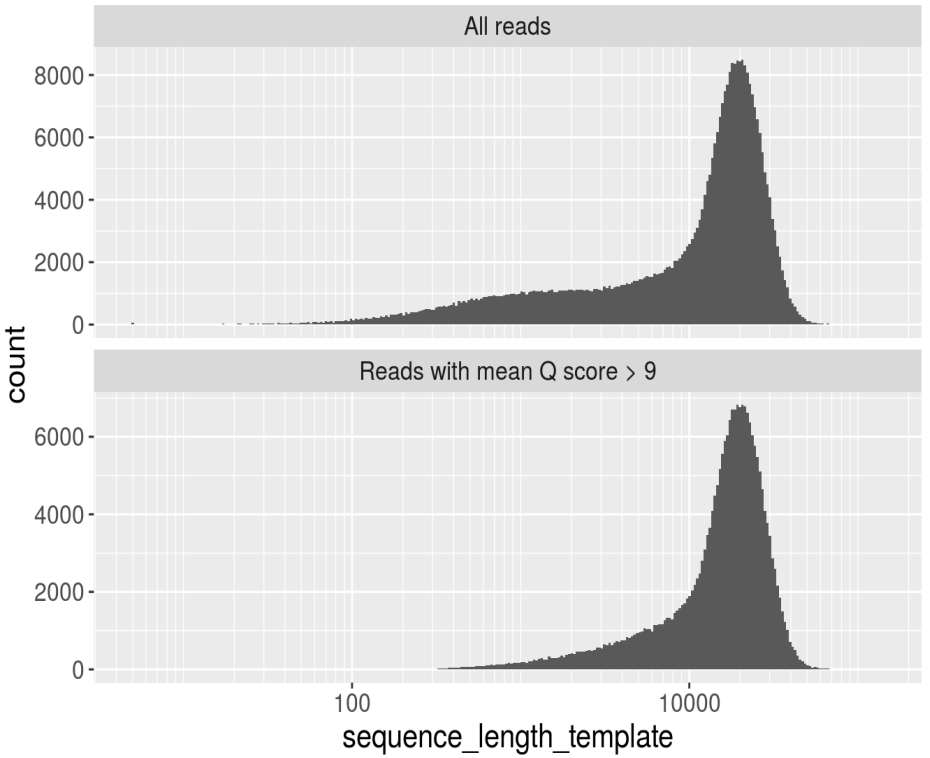
The signal is measured from five bases at a time.
But timing is irregular.

Read length and score filtering

Bad run



Good run

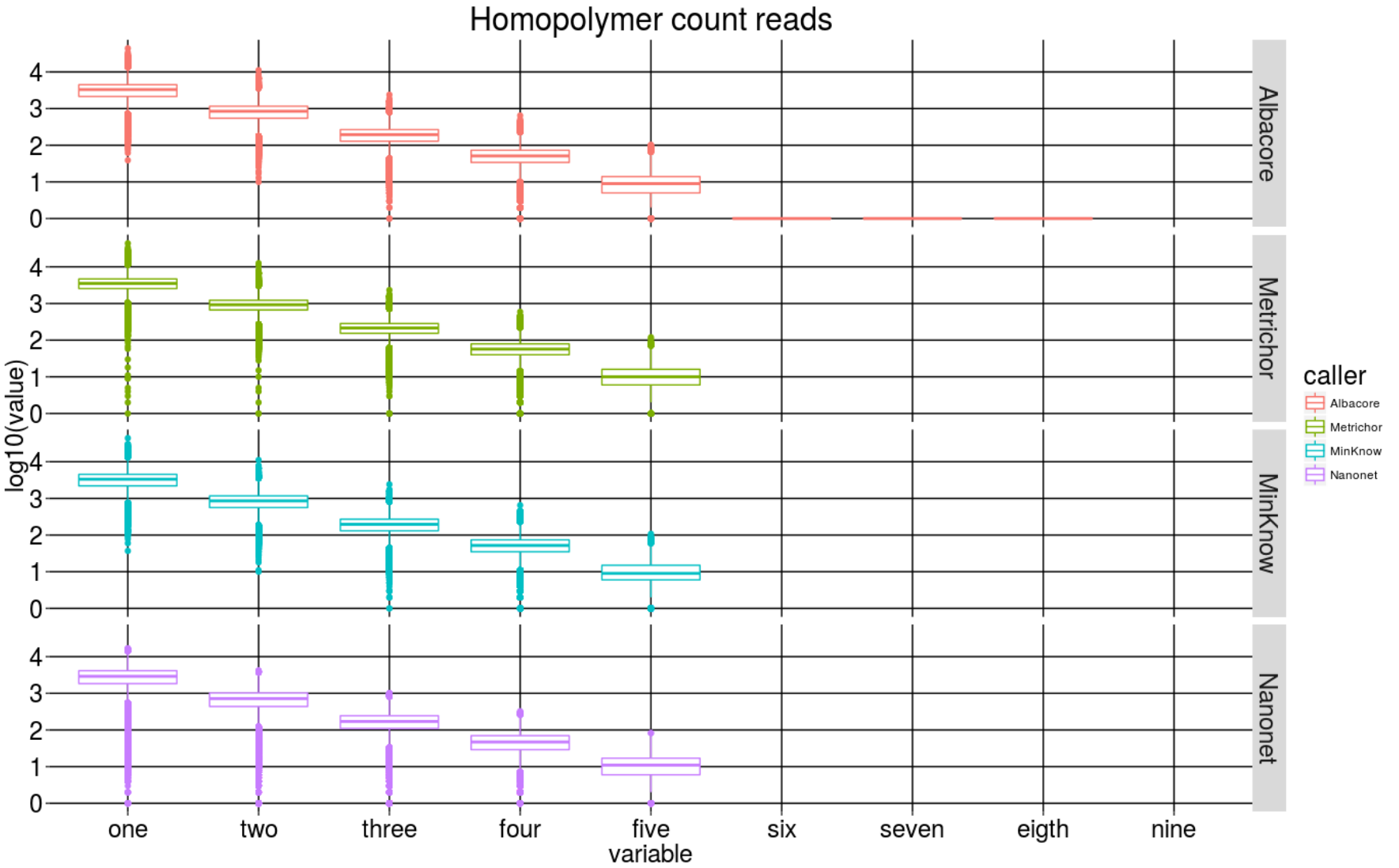


http://www.robertlanfear.com/blog/files/nanopore_performance.html

ONT Read calling, cleaning and filtering

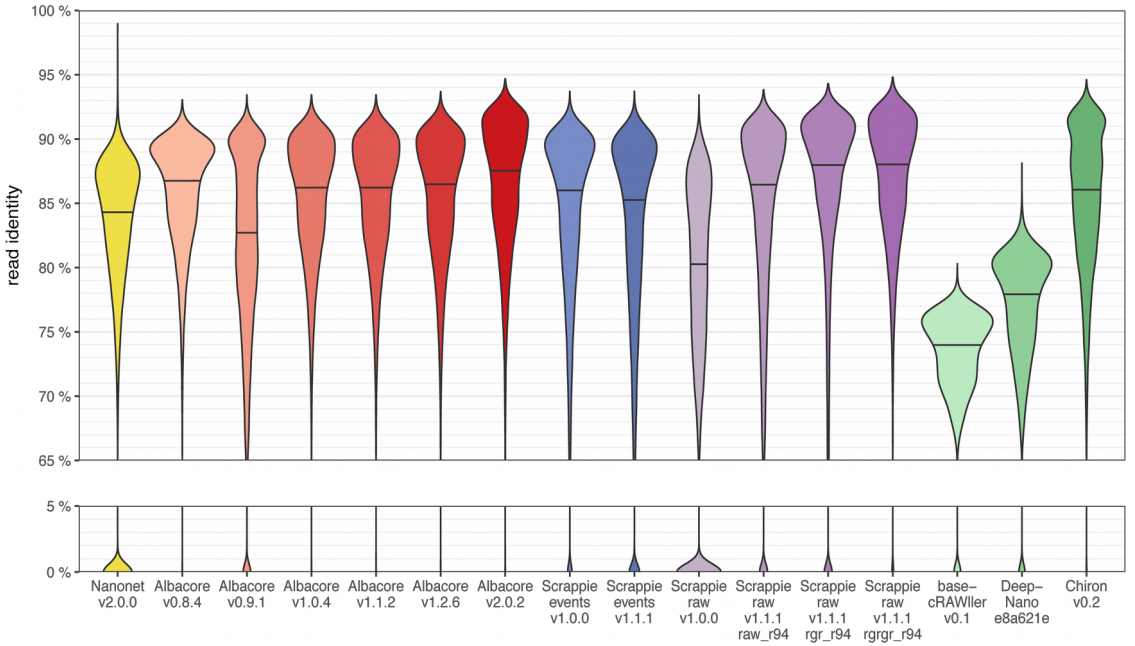
- Sequencer : raw fast5 files
- Step 1 : producing sequence information from fast5
  - **Albacore** : on the cluster or local server
- Step2 : extracting fastq from fast5
  - **Poretools** https://github.com/arq5x/poretools
- Step 3 : quality filtering using the sequencing_summary.txt information
  - Shell script
- Step 4 : removing adapters
  - Porechop https://github.com/rrwick/Porechop

Homopolymers in reads

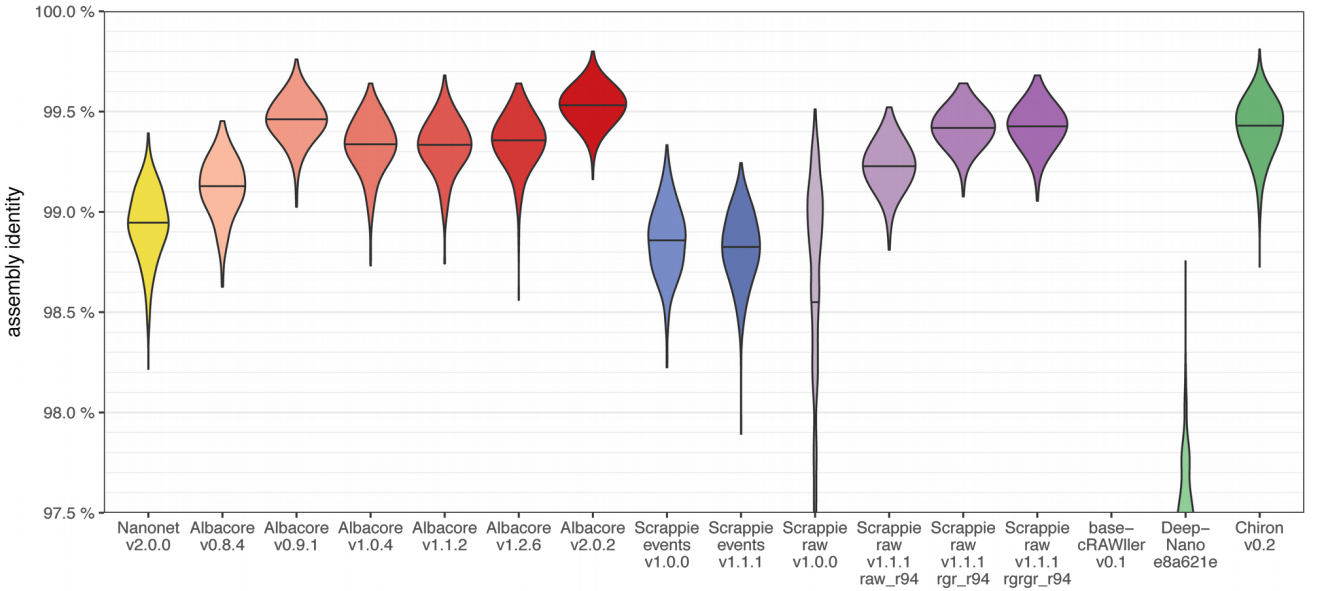


https://github.com/rrwick/Basecalling-comparison

Read identity



Assembly identity



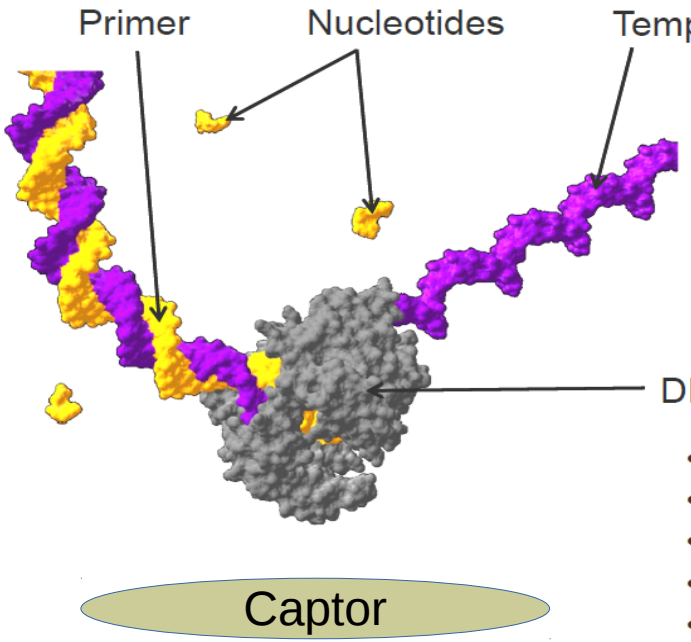
ONT Read analysis conclusions

- Production up to 150Gb per cell for Promethion (15Gb for MinION and GridION).
- The error profile contains :
  - 6% of deletions
  - 4% of substitutions
  - 2% of insertions
- Follow the caller version updates to see if the next version are able to produce the long homopolymers
- Remove adapters with porechop.
- The assembly polishing has still to be done with reads produced by a technology without homopolymer bias.
- Some DNAs are harder to sequence because they do not go easily through the pores.

PacBio technology

SMRT® Cells

A well is called : ZMW
(Zero-mode waveguide)



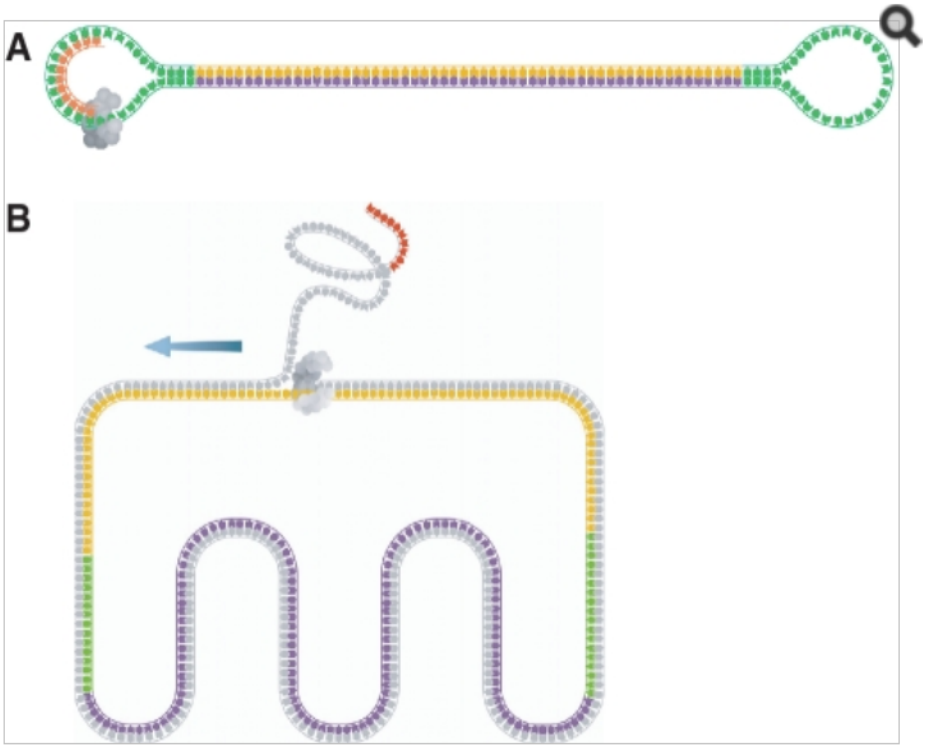
- Fast: hundreds of bases per second
- Processive
- Frugal
- Faithful: 1 in 10^5
- Small ~10nm

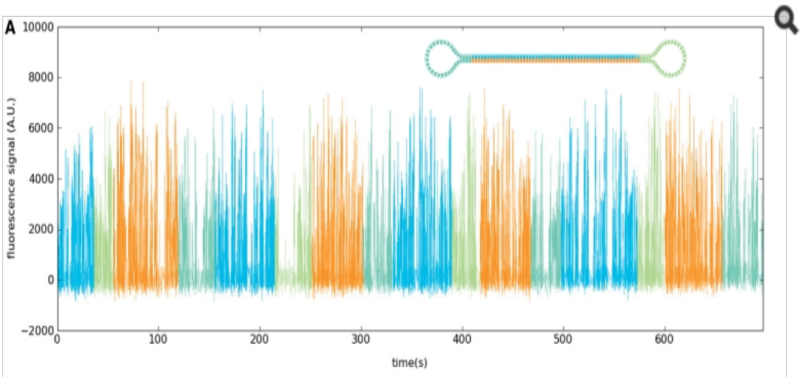
PACIFIC BIOSCIENCES®

Circular consensus sequences

PMC full text: Nucleic Acids Res. Aug 2010; 38(15): e159.
Published online Jun 22, 2010. doi: 10.1093/nar/gkq543

Figure 1.





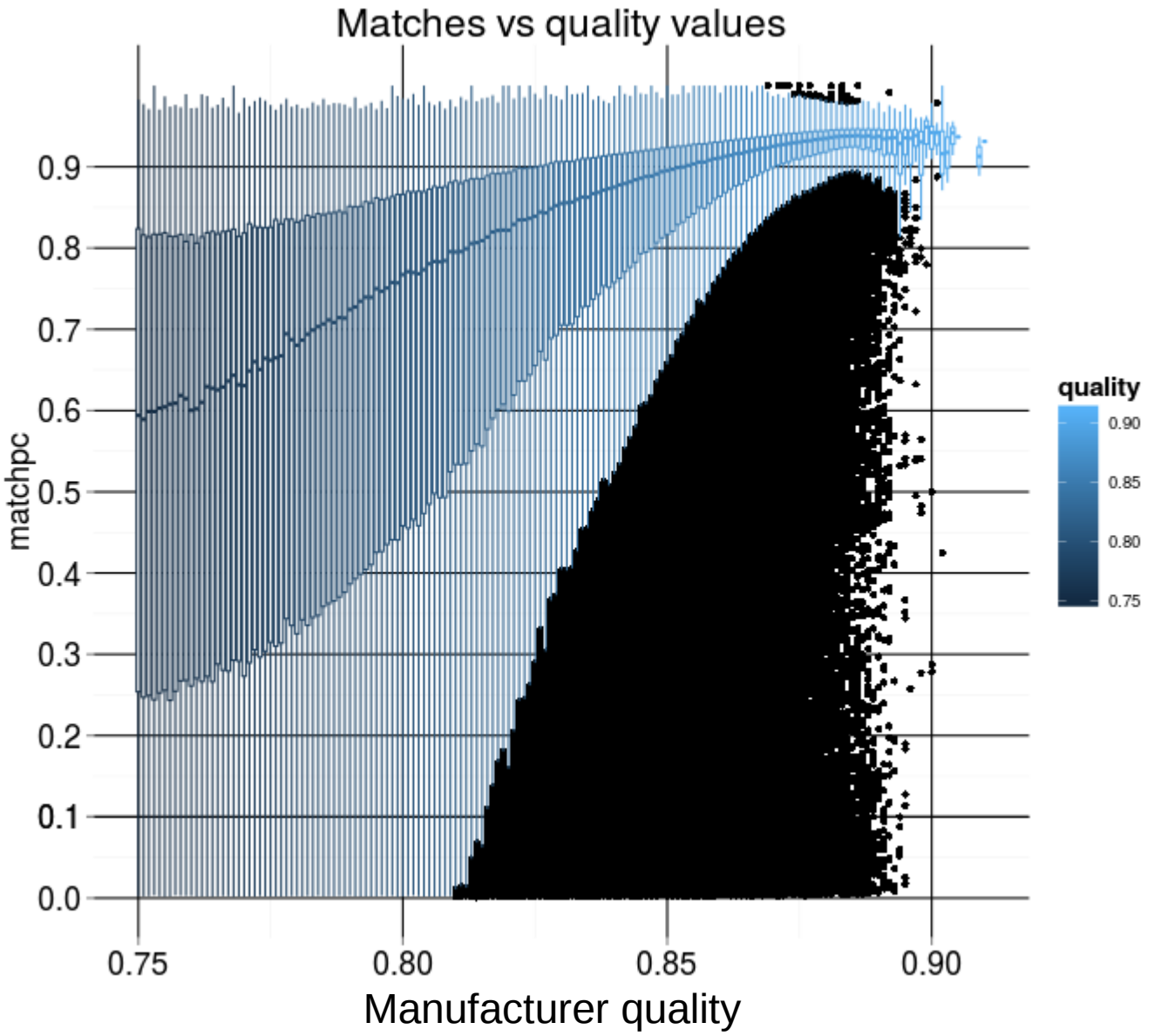
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2926623/

Read calling

- **RSII** : The raw signal is transformed to bas.h5 and bax.h5 files.
- **Sequel** : The raw signal is transformed to bam files.
- Fasta or fastq are extracted using PacBio a software package **bash5tools.py** (for bax files) **bam2fasta** and **bam2fastq** (for bam files)

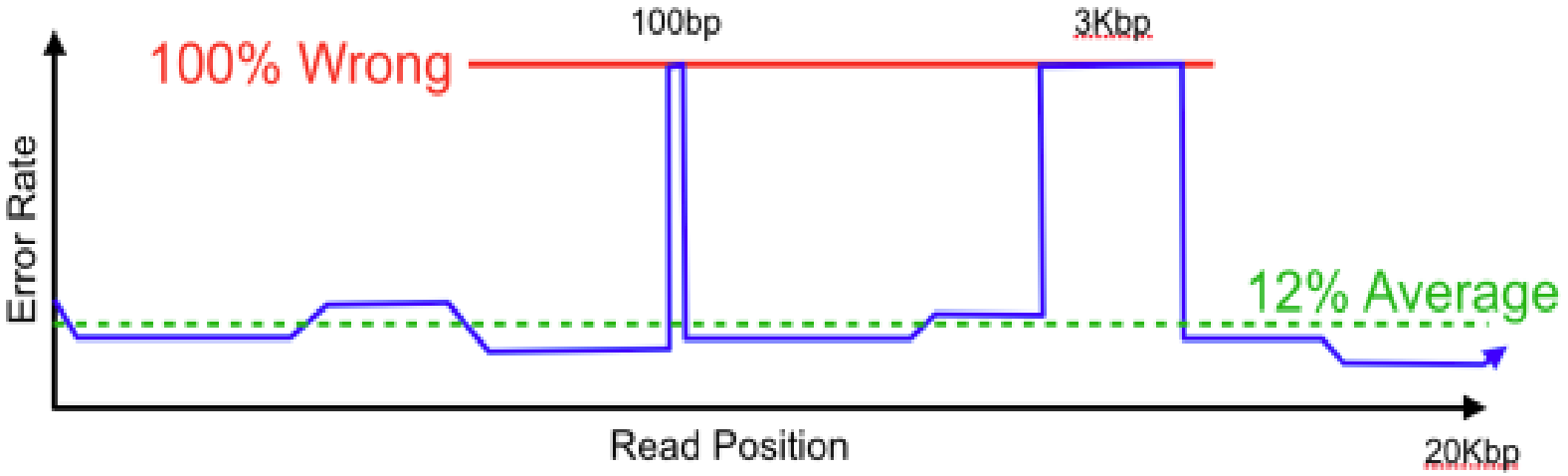
```
@m150619_093250_42174_c100795682550000001823166309091510_s1_p0/109/0_4936 RQ=0.879
@m150619_093250_42174_c100795682550000001823166309091510_s1_p0/109/4981_9942 RQ=0.879
@m150619_093250_42174_c100795682550000001823166309091510_s1_p0/109/9988_10378 RQ=0.879
@m150619_093250_42174_c100795682550000001823166309091510_s1_p0/157/0_7588 RQ=0.871
@m150619_093250_42174_c100795682550000001823166309091510_s1_p0/157/7628_15139 RQ=0.871
@m150619_093250_42174_c100795682550000001823166309091510_s1_p0/157/15186_22778 RQ=0.871
@m150619_093250_42174_c100795682550000001823166309091510_s1_p0/157/22820_30464 RQ=0.871
@m150619_093250_42174_c100795682550000001823166309091510_s1_p0/157/30510_36641 RQ=0.871
```

Read quality versus alignment

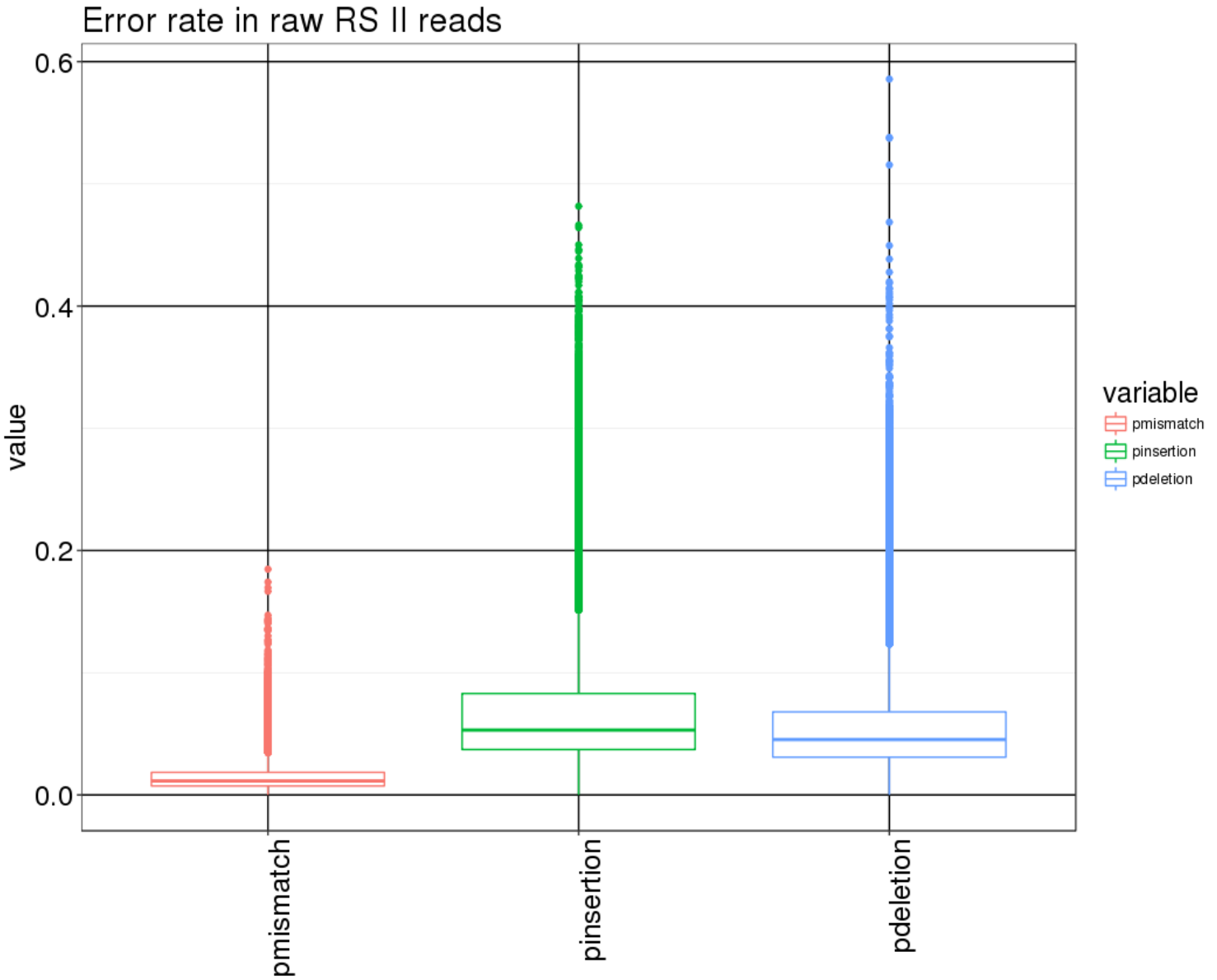


Polymerase errors

https://dazzlerblog.wordpress.com/



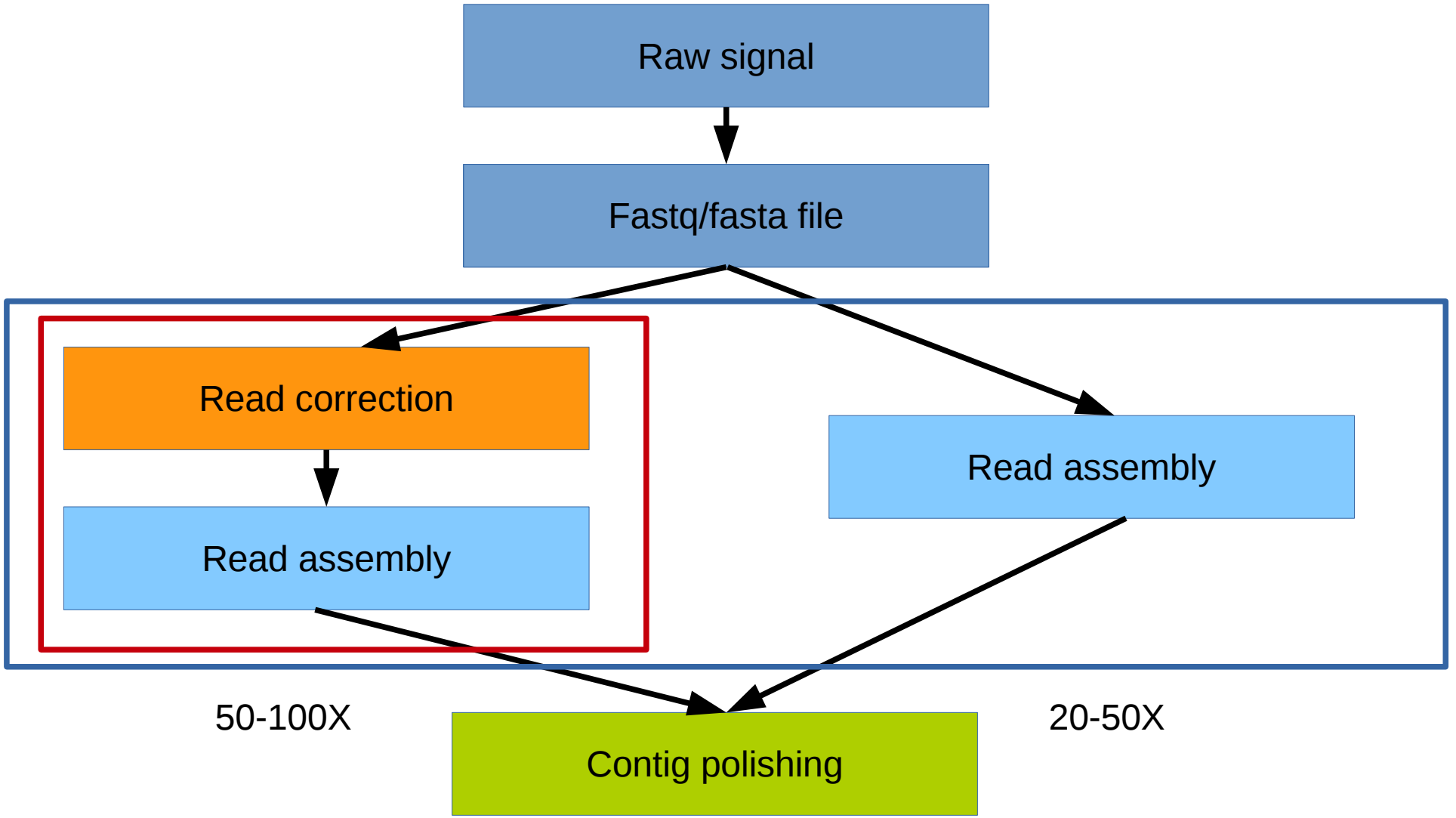
PacBio RSII error profile



PacBio read analysis conclusions

- Sequel produces up to 20 Gb per cell (RS II produced up to 1.5 Gb per cell)
- Only one caller provided for PacBio raw signal
- Random errors
- Random selection of the reads along the genome
- But some parts of the genomes are still not sequenced (chicken PB assembly is missing micro chromosomes)

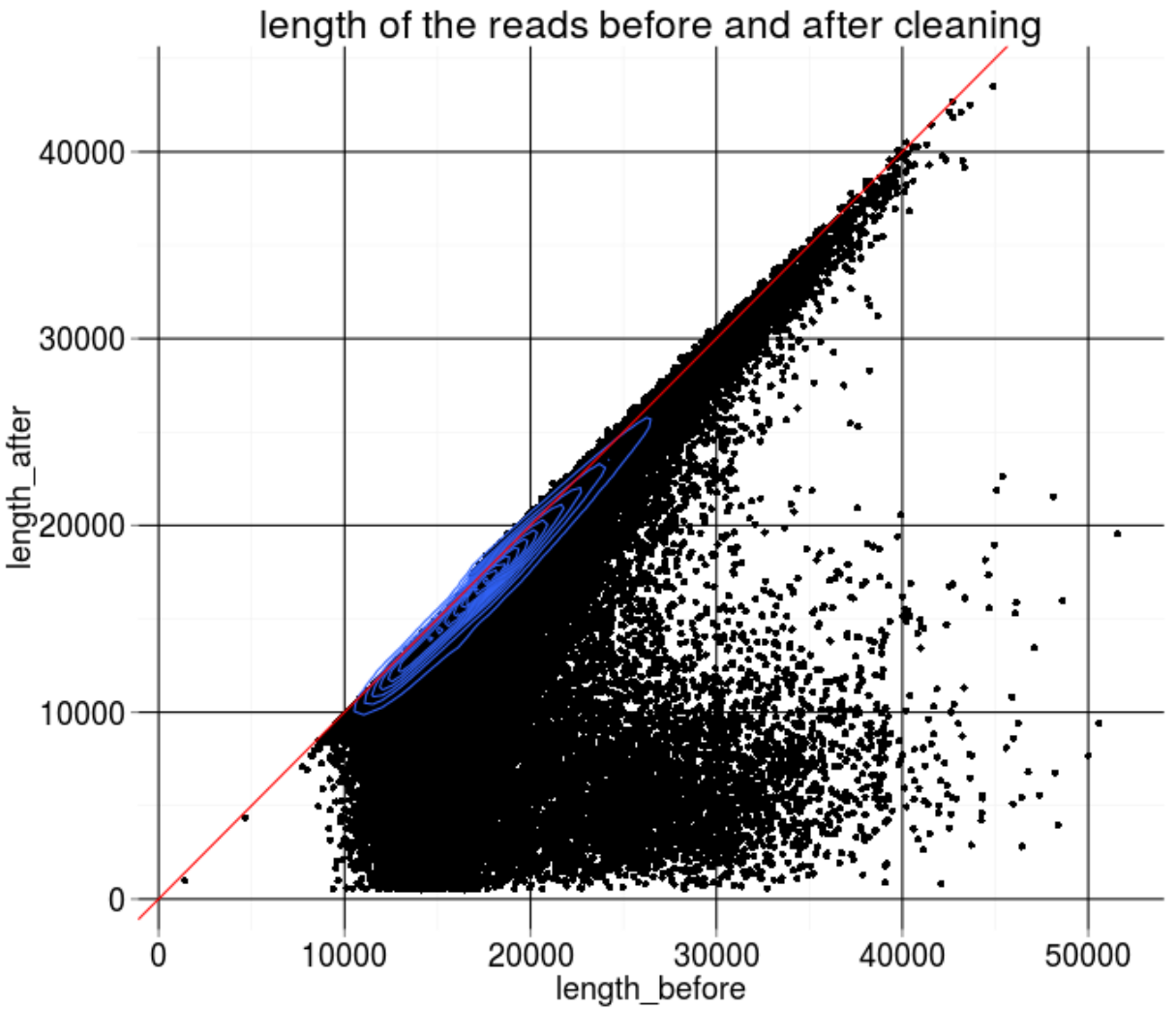
Assemblers with read correction



Read correction process

- Correction strategies
  - External reads : Illumina
  - Internal reads :
    - Long reads corrected by short ones
    - All versus all
- Correction pipeline
  - Read alignment
  - Consensus calling

Corrected versus raw lengths

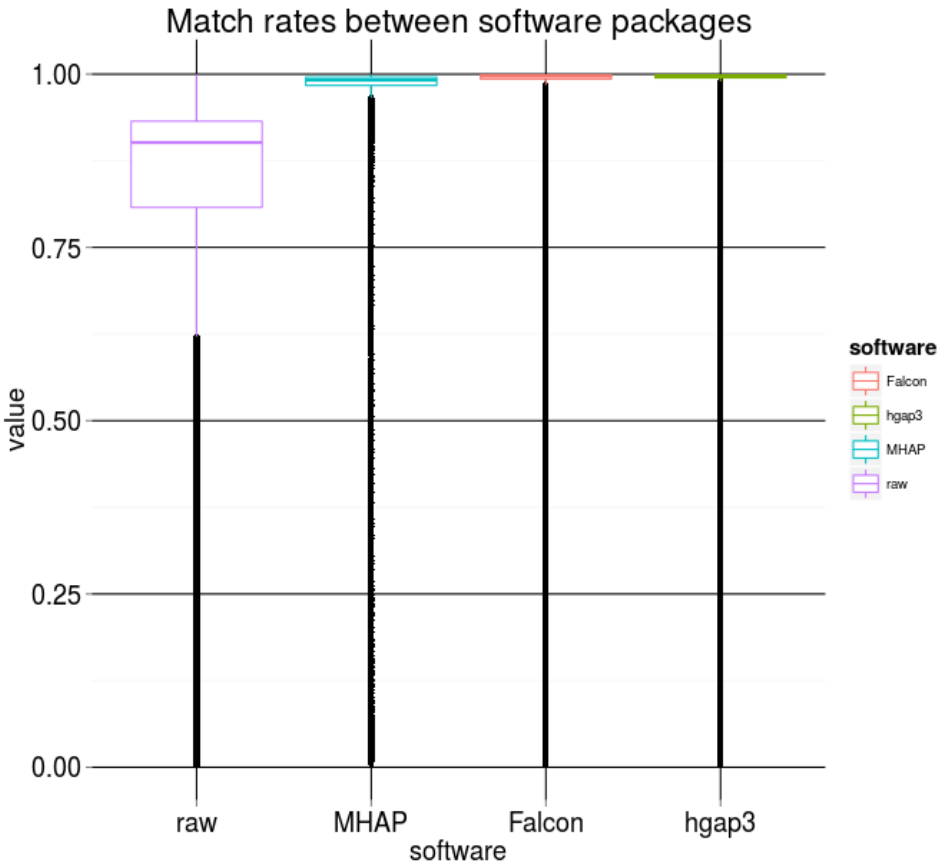


215,843read pairs : before and after

Assemblers with read correction

	HGAP	CANU	Falcon
Read correction **aligner**	Read splitting for correction. **blasr**	**MHAP**	1 read per film selection **daligner**
Selection		40X of longest reads	Size selection criteria
Assembly of corrected reads	First steps of WGS-assembler + Specific module	WGS-assembler	First steps of WGS-assembler + Specific module

Correction impact : PacBio



Falcon

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.004315 0.992900 0.996900 0.969200 0.998600 1.000000
```

hgap3

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.003802 0.994700 0.997100 0.969100 0.998300 1.000000
```

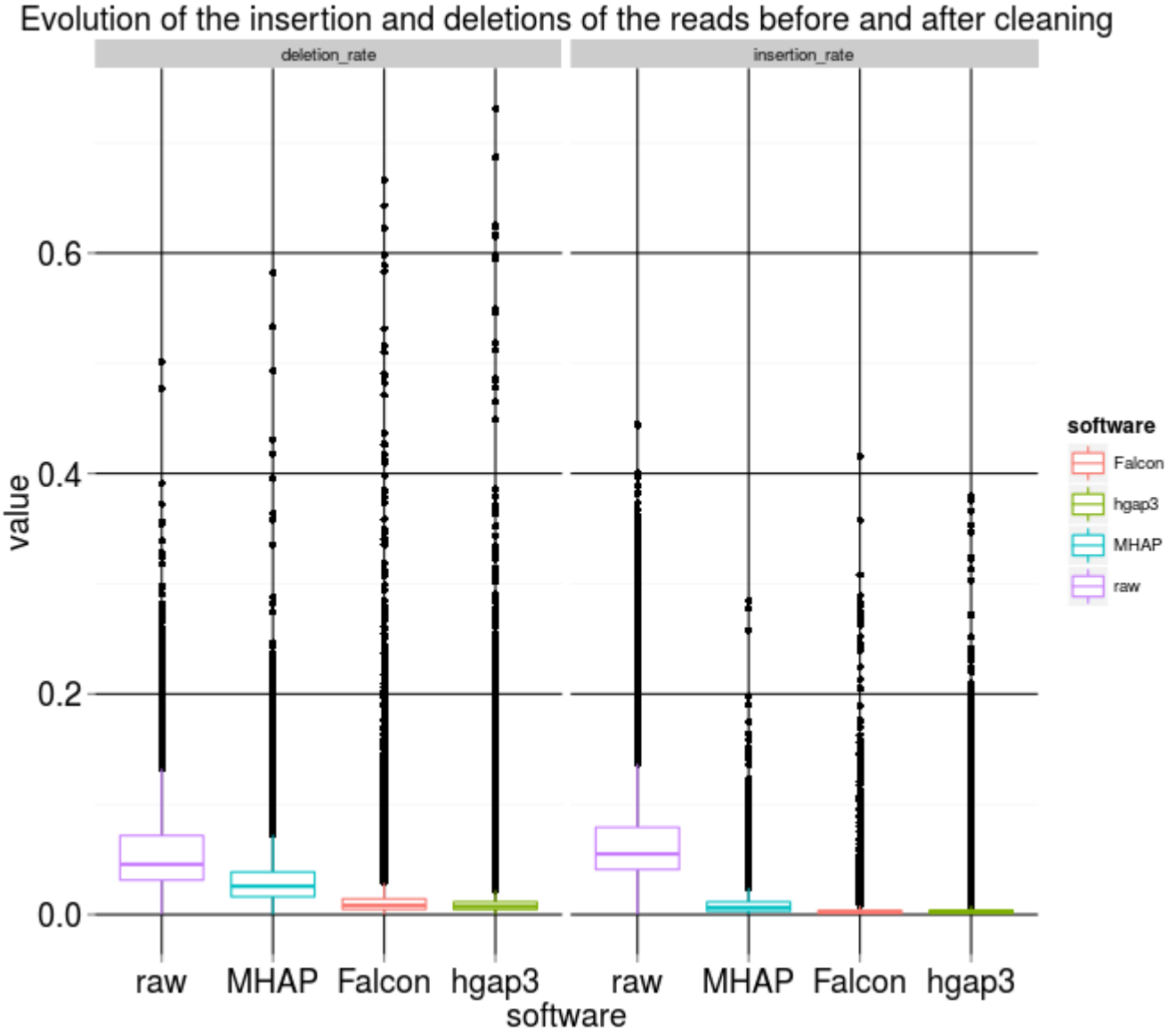
MHAP

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.005378 0.983600 0.991300 0.966600 0.995600 1.000000
```

raw

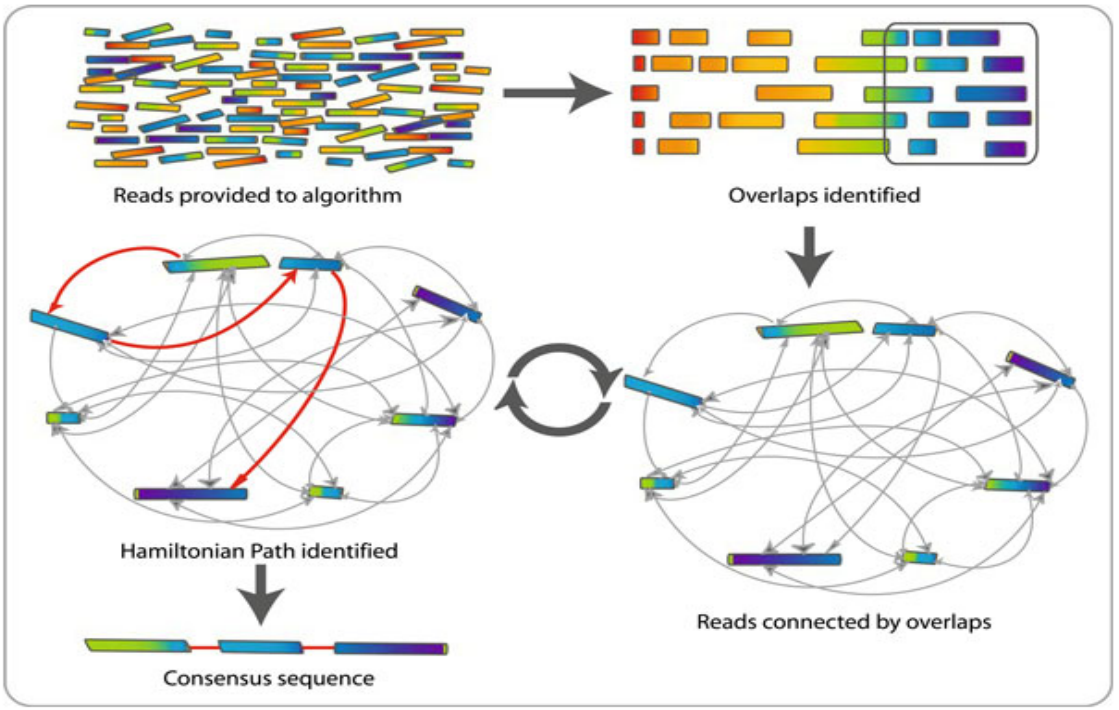
```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.002205 0.807700 0.901200 0.810400 0.932000 1.000000
```

Insertion and deletion rates



Assembling corrected reads

- OLC (Overlap Layout Consensus) assembler



http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3055744/

- Celera assembler used by hgap, Canu, Falcon

Contig assembly metrics

Species **Assembler**	Genome size	Coverage	Total contig size	Contig count	N50	L50
Caenorhabditis elegans **Falcon PB**	100 Mb	47X	101,72 Mb	122	2,022,653	17
Aeschynomene evenia **Hgap PB**	400 Mb	120X	374.45 Mb	5,711	648,407	87
Ictalurus punctatus **Falcon PB**	900 Mb	120X	826,04 Mb	1 ,554	4,431,159	50
Oryzias javanicus **Falcon PB**	900 Mb	50X	865,44Mb	1,286	3,821 ;811	59
Arabidopsis thaliana **Canu ONT**	120 Mb	81X	121,09 Mb	197	16,062,269	5

Assemblers without read correction

- Miniasm, Smartdenovo and wtdbg are members of this “new” family
- Improves speed (no correction).
- Can work with less read depth.
- They can also assemble corrected reads.

Contig assembly metrics

Species **Assembler**	Genome size	Coverage	Total contig size	Contig count	N50	L50
Caenorhabditis elegans **Miniasm PB**	100 Mb	47X	106,81 Mb	129	2,214,883	17
Oryzias javanicus **Miniasm PB**	900 Mb	50X	816,95 Mb	2,638	685,722	323

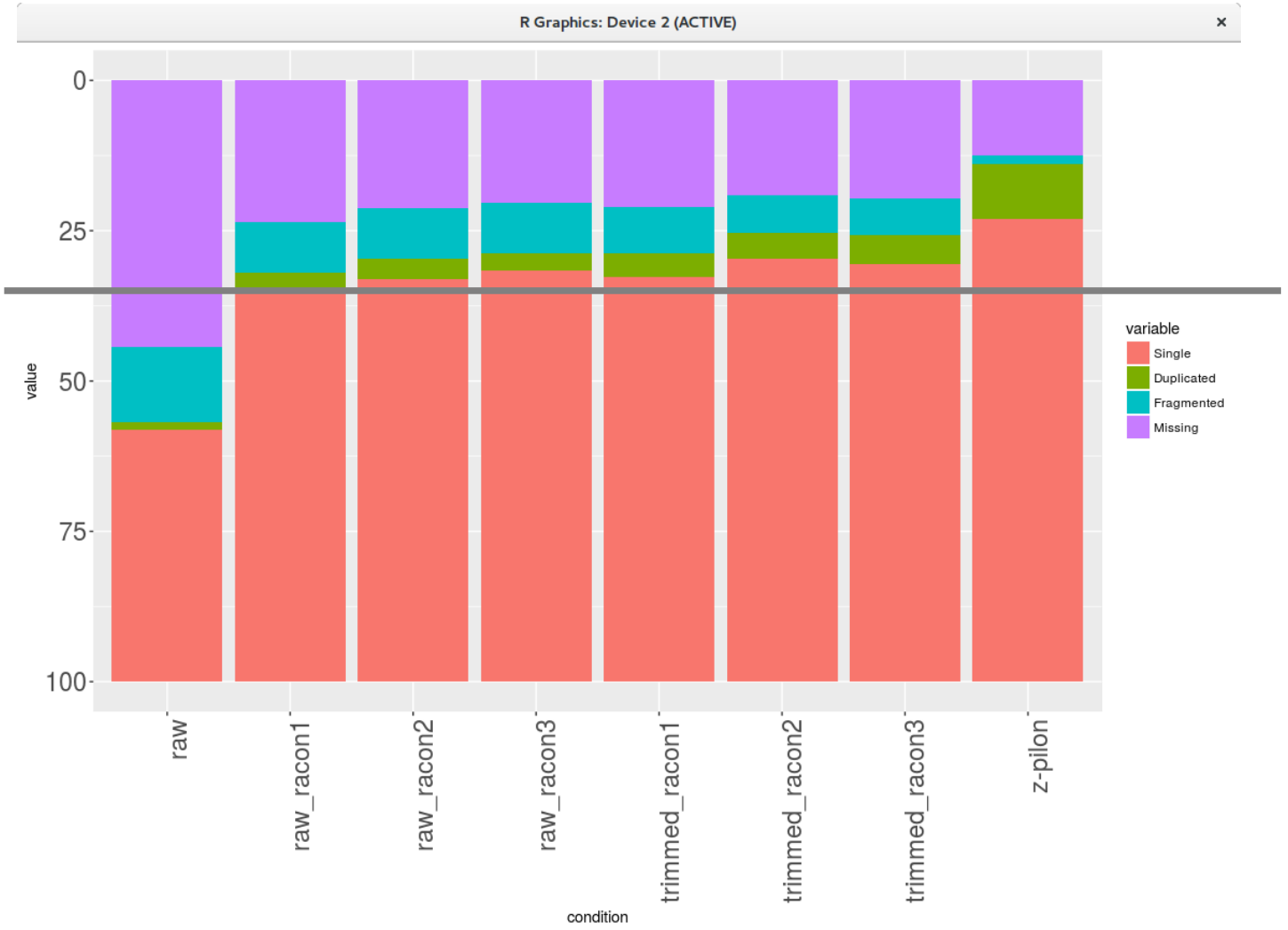
Assembly polishing

- There are still errors in the assembled scaffolds (mainly insertions/deletions) and a lot with miniasm and smartdenovo!
- **Quiver/Arrow :** is an algorithm for calling highly accurate consensus from multiple **PacBio** reads, using a pair-HMM exploiting both the basecalls and QV metrics to infer the true underlying DNA sequence.
- **Nanopolish** : is using the **ONT** fast5 files to create the fastq which is the aligned with bwa and the resulting bam is used to polish the genome.
- **Pilon :** is using a short (or long) good quality read alignments to correct the consensus sequence.
- **Racon** : is using long reads (fastq) and overlaps to correct the consensus sequence.

Polishing impact

BUSCO v3

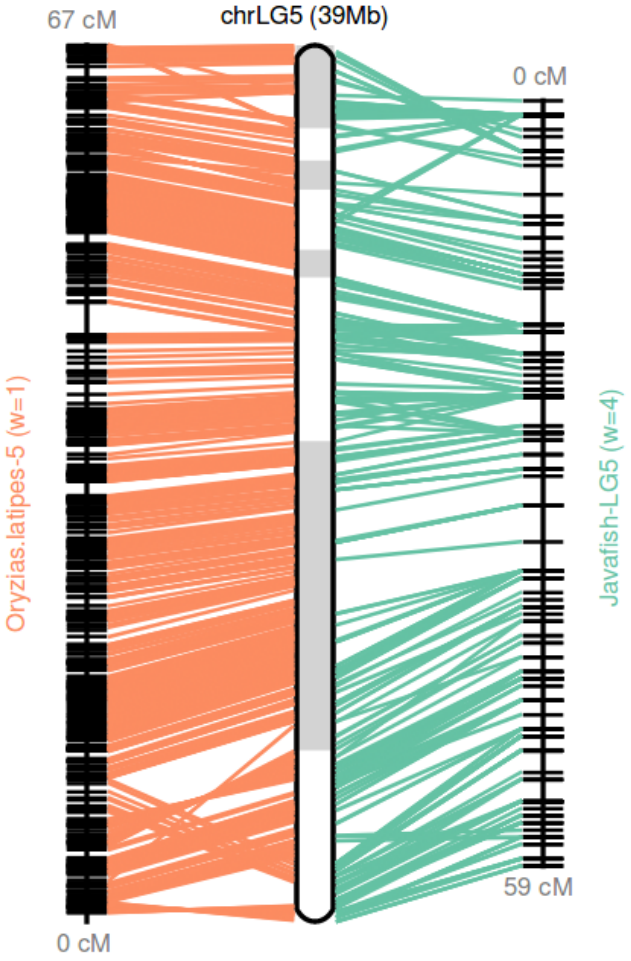
https://busco.ezlab.org/

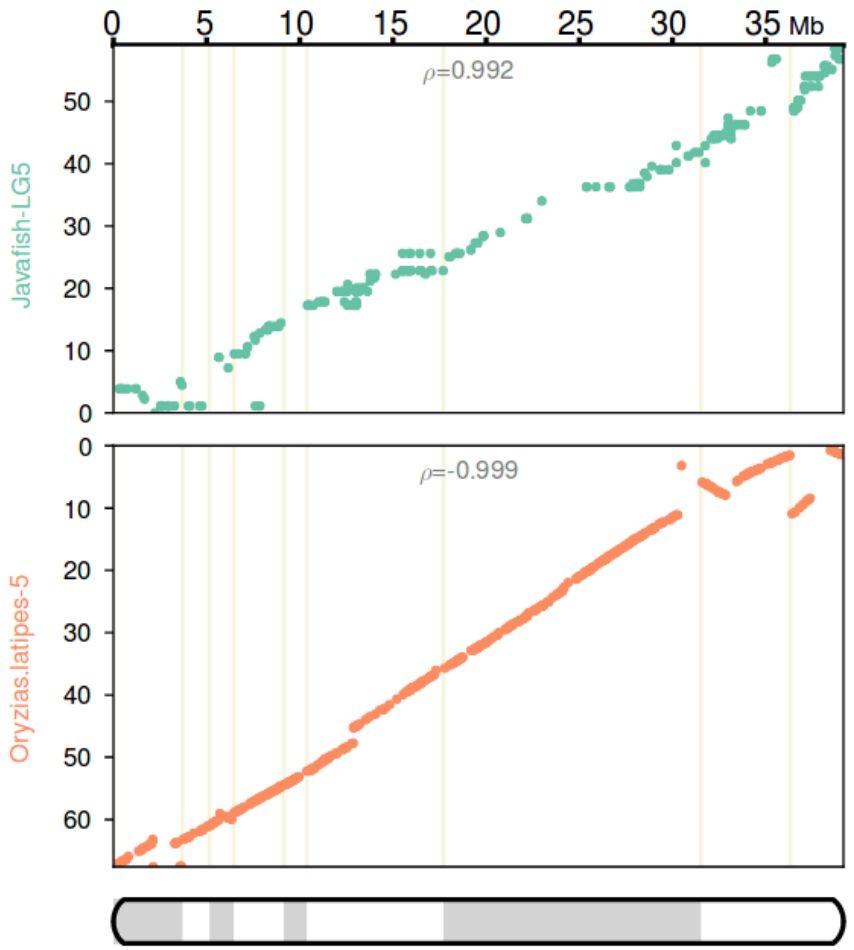


From contigs to chromosomes

- **Optical mapping** : fluorescent marking of restriction sites of very long DNA molecules (up to Mb) to extract signature used to bridge contigs having these signatures.
- **10x chromium** : shallow tagged sequencing of very long DNA fragments with Illumina machines. Read alignments enable scaffolding.
- **Genetic map** : marker assisted contig bridging
- **HiC** : chromosomal interaction sequencing gives the contig order on the chromosomes.

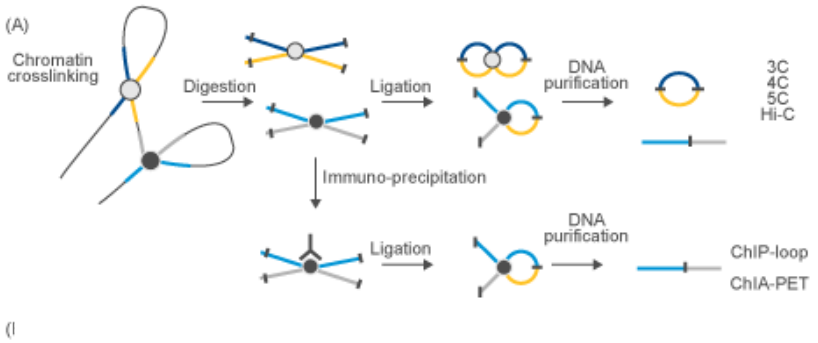
Building chromosomes : genetic map

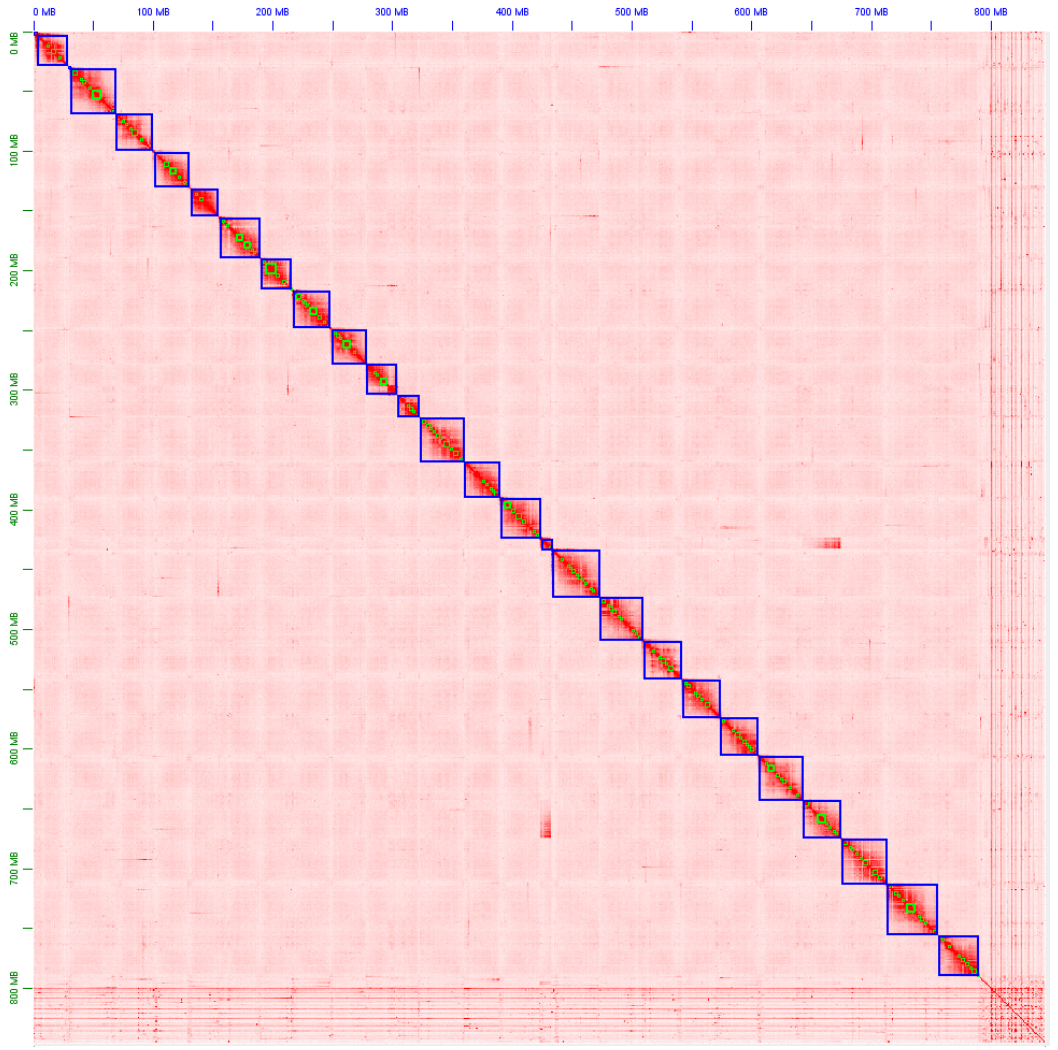




Building chromosomes : Hi-C

Hi-C protocol



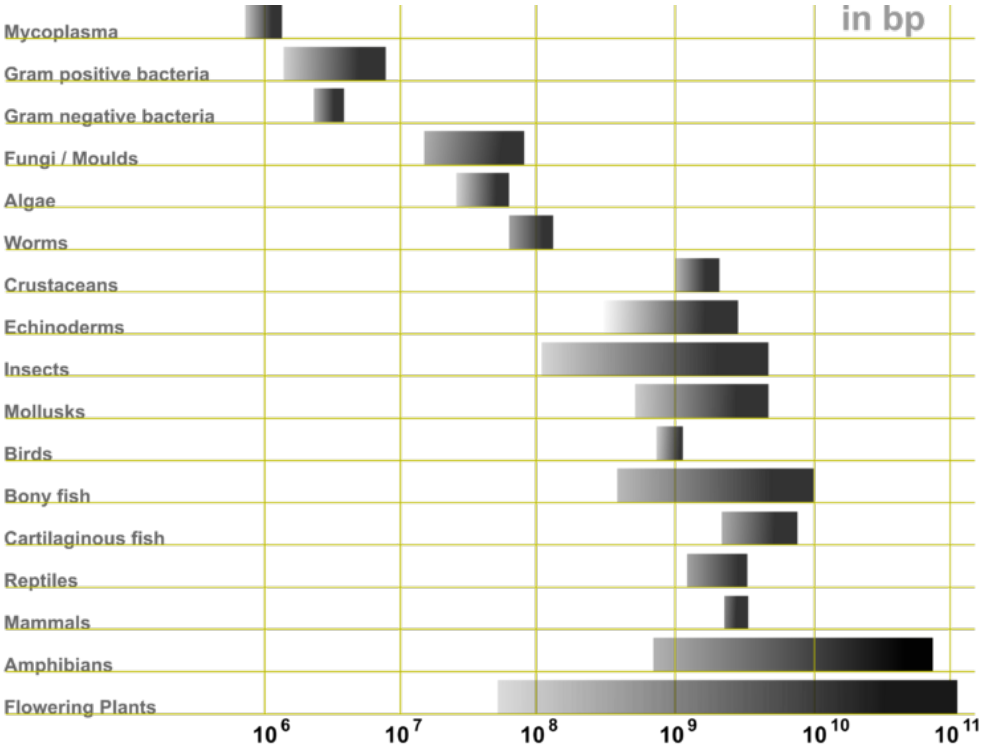


Perca flavescens

http://www.abcam.com/protocols/getting-started-with-chromatin-conformation-capture-3c

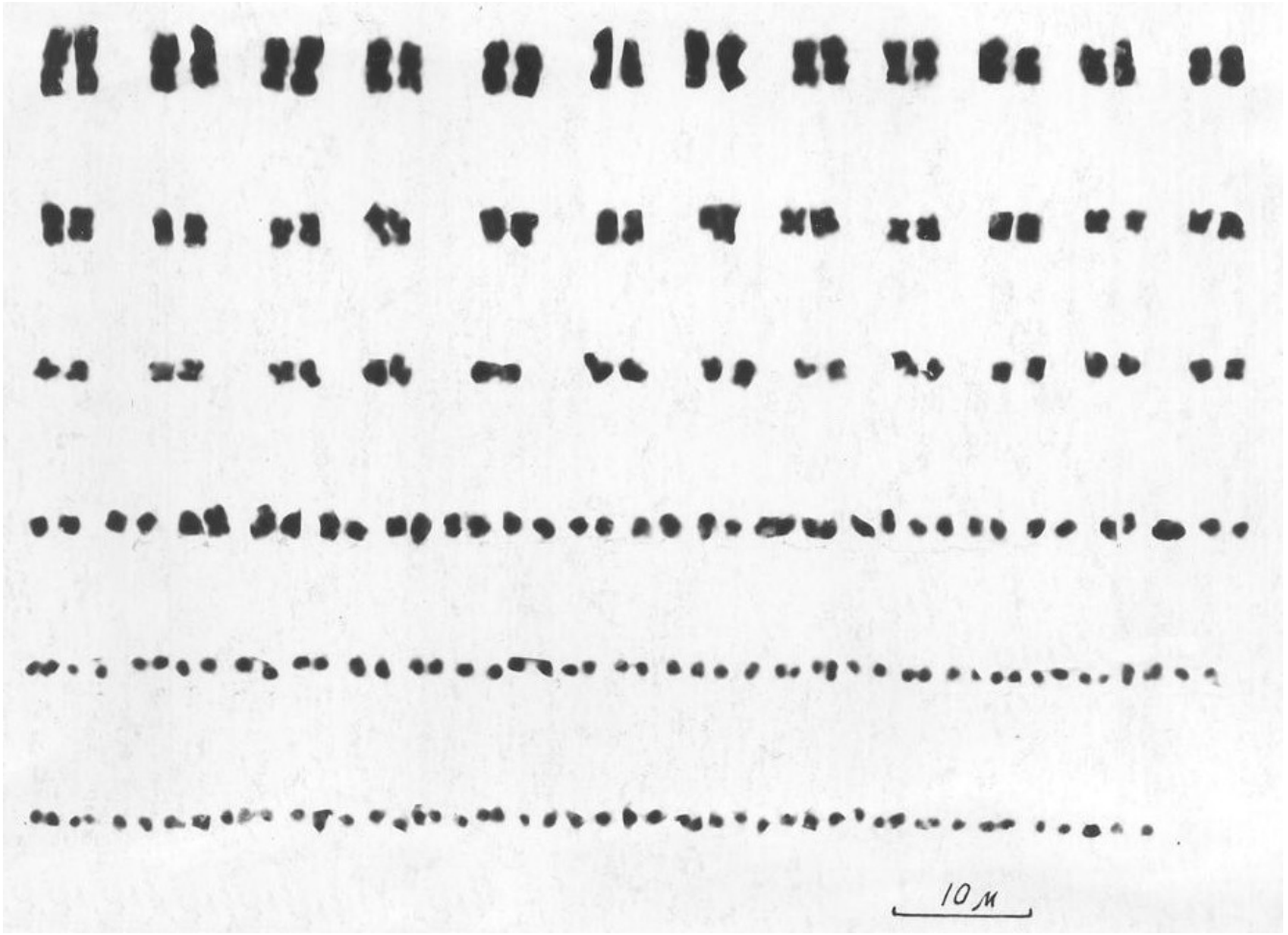
What makes genomes difficult to assemble ?

- Genome size
- Number of chromosomes
- Repeat content
- Repeat size (structure)
- Heterozygocity
- Ploïdy
- DNA conformation
- Contamination



https://en.wikipedia.org/wiki/Genome_size

Sturgeon karyotype



http://www.unife.it/dipartimento/biologia-evoluzione/progetti/geneweb/immagini/babur73_g.jpg

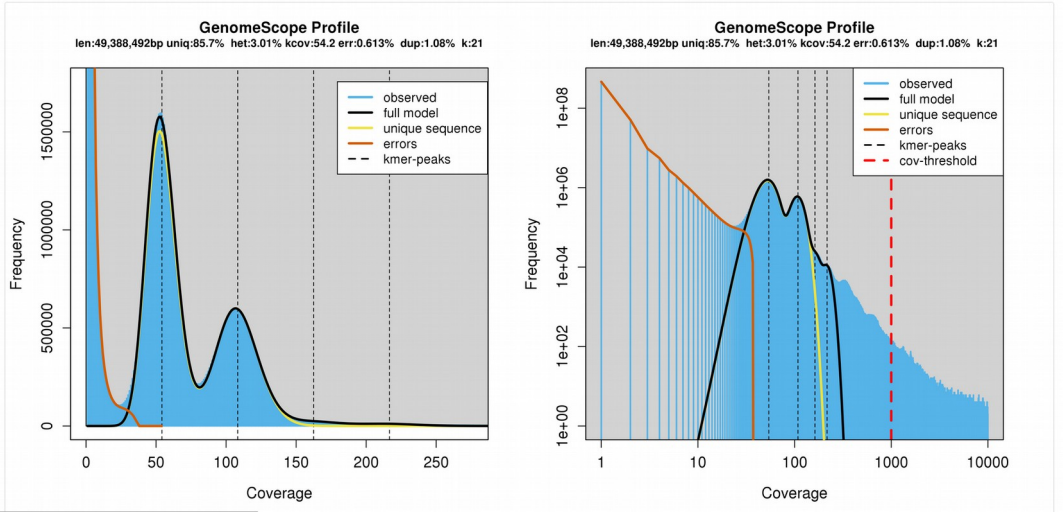
2n ~ 209...249

Difficult to assemble genomes 1

Polyploid genomes

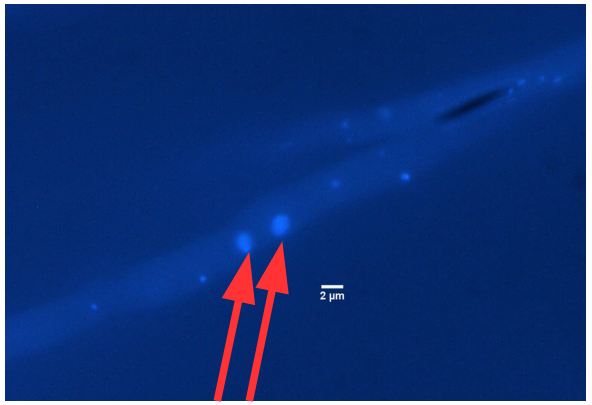
Ganoderma fungi

dikaryotic stage



	MiSeq	PacBio
software	**DiscovarDeNovo**	**CANU**
version	**52488**	**1.5**
Number of contigs	46,450	915
Number of contigs not in scaffolds	46,250	915
Total size of contigs	84,459,515	84,137,414
N50 contig length	9,423	166,220
L50 contig count	2,172	136

Haplomerging



dikaryotic stage

Results

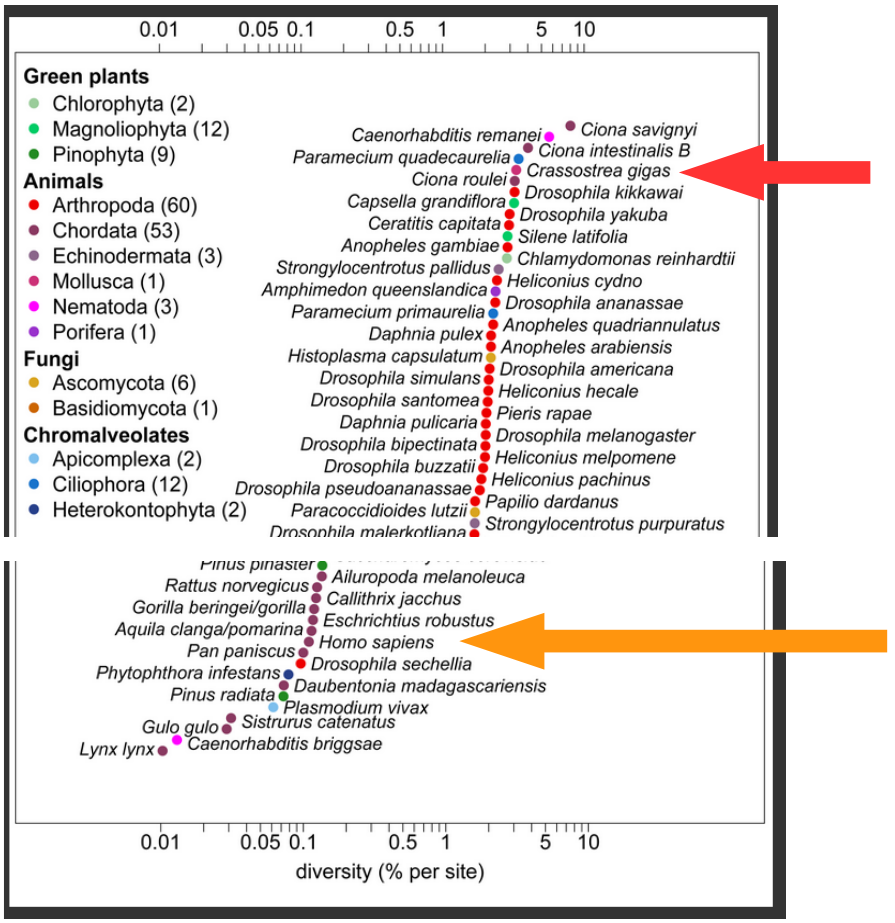
GenomeScope version 1.0
k = 21

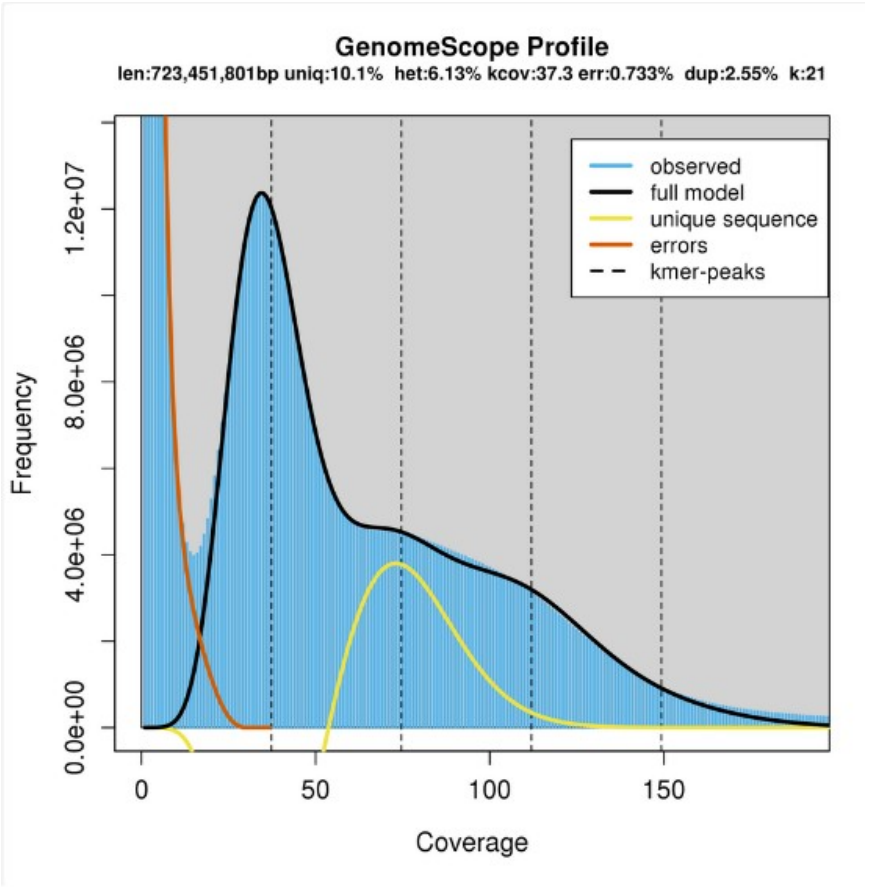
property	min	max
Heterozygosity	3.0009%	3.02719%
Genome Haploid Length	49,334,013 bp	49,388,492 bp
Genome Repeat Length	7,070,219 bp	7,078,026 bp
Genome Unique Length	42,263,794 bp	42,310,465 bp
Model Fit	96.393%	97.6122%
Read Error Rate	0.612984%	0.612984%

	PacBio + Nanopore	PacBio + Nanopore	PacBio + Nanopore
software	**CANU**	**haplomerger**	**haplomerger**
version	**1.5**	**2.0**	**2.0**
Number of contigs	436	218	218
Number of contigs not in scaffolds	436	218	218
Total size of contigs	76,821,474	56,529,789	49,372,845
N50 contig length	323,663	**1,068,631**	**889,398**
L50 contig count	55	**17**	**18**

Difficult to assemble genomes 2

Heterozygosity: oyster





http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1001388

Oyster assembly metrics

```
GenomeScope version 1.0
k = 21
property                  min               max
Heterozygosity            6.09571%          6.17102%
Genome Haploid Length     720,368,835 bp    723,451,801 bp
Genome Repeat Length      647,690,896 bp    650,462,822 bp
Genome Unique Length      72,677,940 bp     72,988,980 bp
Model Fit                 93.522%           97.9978%
Read Error Rate           0.733165%         0.733165%
```

metrics	CANU PacBio
Number of contigs	19,188
Total size of contigs	1,307,520,554
Longest contig	2,388,545
Shortest contig	1,158
Mean contig size	68,143
Median contig size	45,702
N50 contig length	86,828
L50 contig count	3779

Conclusions

- DNA quality (fragment length) has a direct impact on read length. You have to find the best fragment length given the sequencing technology.
- We can assemble small to large genomes with PacBio or Nanopore reads.
- Polishing is mandatory!
- There are still genomes which are difficult to assemble.

Acknowledgments

Philippe LeLeux (IRD)

Alain Vignal (INRA)

Jean-François Arrighi (IRD)

Jérôme Gouzy (INRA)

Sylvie Quiniou (USDA)

Roberto Lleras (PacBio)

Cécile Donnadieu (INRA)

Yann Guiguen (INRA)

Alain Roulet (INRA)

Get-Plage team members