



# 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.



GeT\_Genotoul

@GeT\_Genotoul

Replies to [@CurtisKapsak](#) [@Apple](#) [@illumina](#)

You are right, it was a part of the **#PromethION** from [@nanopore](#) which arrived today. The **#GridION** arrived one year ago : [twitter.com/GeT\\_Genotoul/s...](https://twitter.com/GeT_Genotoul/status/1000000000000000000)

[Translate Tweet](#)



9:59pm · 6 Sep 2018 · Twitter Web Client

# 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

Sbjct	Query	Sequence	Length
3134459	212	CTGTT-AAAGGCCACATATTGTTAATAATTTTATGAAATAAAATTAAATAC-A	3134516
		aaaaataagaatatgttcttaatattattatacgtaAGAAAGTACattacttataat	271
3134517		AAATAAGAATATGTCCTTAATATTATTCATAACGCAAGAAAGTACATTACTATTAAAT	3134576
	272	tattaattactaaaataataatctatTTaaaataacttaacaatattatggaaagst	331
3134577		TATTAATTACTTAAATAATAATCTATTT-AAAATCTAACATAATTATGGGAAGGT	3134635
	332	aaaaacccaaaaatacaatttatctttataactAAATATG-GCATTAAAtccatattt	390
3134636		AAAAACCAAATAACATTTCTTATAAACTAAATAATGTACATTAAATCCATATTG	3134695
	391	aattt-aaaagacAAATCGGGATACATAGAACatattaaatttgacTATGTT--TACA	447
3134696		AATTAAAAAGACAAATCGGGATACATAGAACATTTATTTATGGACTATGTTTATACA	3134755
	448	AGTTGTTTgcattattgtatTTgaaattaaaa-ttaata-ttttcttaacatat	505
3134756		AGTTGTTTGCATTATTGATAATTGAAATTAAAAATTAAATTTTCTTAACATAT	3134815
	506	acatatagttGCATGTATGTGATAACAGAACATCCATACATGTACATATAAGTTACTAT	565
3134816		ACATATAGTTGCATGTATGTGATAACAGAACATCCATACATGTACATATAAGTTACTAT	3134875
	566	GTACAGAAagacattt-tttttagtGTAATATACTTCGCCTTATCCTATATCTAT	624
3134876		GTACAGAAAGACATTATTTTATGAGTGTAAATAATCC-TGCCTTATCCTATATCTAT	3134934
	625	TACCCATGGCTCTTACTTTGqaataaattactattaataatataatccatGTGAGC	684
3134935		TACCCATGGCTCTTACTTTGGAATAAAATTACTTAAATAATATAATCCATGTGAGC	3134994
	685	AACTTGTTGATATGtcataaaat-aacaaatgttctcaattttataatttttgc	743
3134995		AACTGTTGATATGTCATAAAATAACAAATGTTCTCAATTATTATAATTGTTTG	3135054
	744	agaaCAAAGTATCATTATTC-TGTGCACATggcataatttt-aattgaa-ttttata	800
3135055		AGAACAAAGTATCATTATTCTTGTGCACATGGCATATTTTAAATTGAATTTTATA	3135114
	801	attaagtgaaactaacatttatatacatatataatataatataatcatgtataat	860
3135115		ATTAAGTAAAATCAACATTATACATATATATATACATAATATACATGTATAATGCT	3135174
	861	ATGTAATGTTAATCGCTTGTATACATTGTCATCACTCAATCATCATCTGCAC-	919
3135175		ATGTAATGTTAATCGCTTGTATACATTGTCATCACTCAATCATCATCTGCACACA	3135234
	920	AATACATTTCCCACCATGTAATAAccataataatttcaatggattttgaatgatga	979
3135235		AATACATTTCCCACCATGTAATAACCAATAATAATTCAATAGTATTGAAATGATGA	3135294
	980	aataaacaaataaagtaTACAGATATATGCcacattttatctattactgTTACTTCATT	1039

# 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

The screenshot shows the header of the Nature Genetics website. The logo 'nature genetics' is on the left. To the right is a decorative background image of a DNA double helix. Below the logo is a horizontal navigation bar with links: Home, Current issue, Comment, Research, Archive ▾, Authors & referees ▾, and 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



[PDF](#)



[Citation](#)



[Reprints](#)



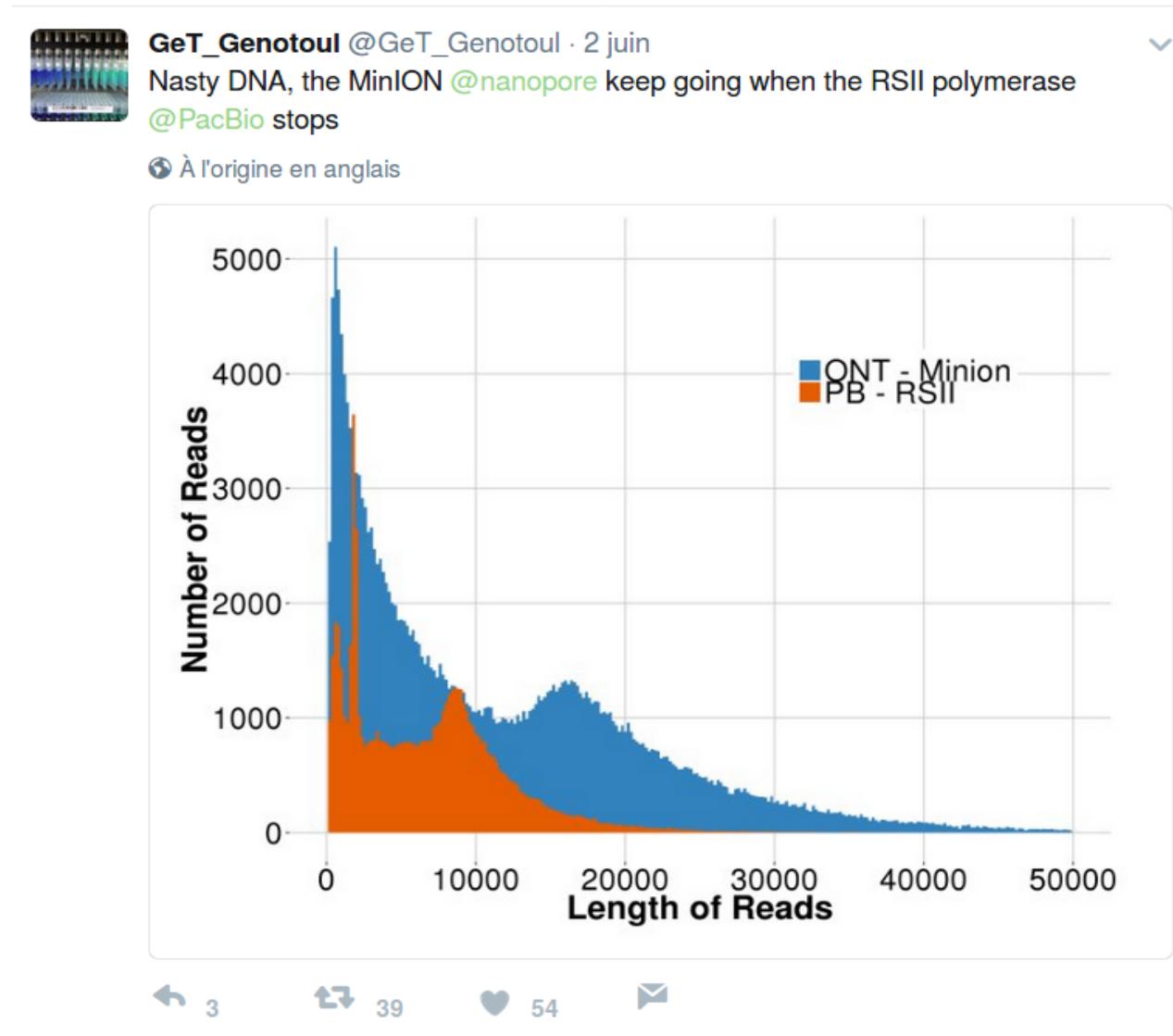
[Rights & permissions](#)



[Article metrics](#)

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/>

The screenshot shows the homepage of the Genome 10K project. At the top, there's a navigation bar with links for My UCSC, People, Calendars, A-Z Index, and a DONATE button. Below the navigation is a logo for the University of California Santa Cruz Genomics Institute and the Genome 10K project. The main content area features a large text block about the project's goal to assemble a genomic zoo of 10,000 vertebrate species. To the right of the text is a photograph of three men standing next to a large mammal skeleton. Below this is another photograph of a group of people standing in front of a building.

The Genome 10K project aims to assemble a genomic zoo—a collection of DNA sequences representing the genomes of 10,000 vertebrate species, approximately one for every vertebrate genus. The trajectory of cost reduction in DNA sequencing suggests that this project will be feasible within a few years. Capturing the genetic diversity of vertebrate species would create an unprecedented resource for the life sciences and for worldwide conservation efforts.

The growing Genome 10K Community of Scientists (G10KCOS), made up of leading scientists representing major zoos, museums, research centers, and universities around the world, is dedicated to coordinating efforts in tissue specimen collection that will lay the groundwork for a large-scale sequencing and analysis project.

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

The screenshot shows the submission page for the 10KP database. At the top, there's a navigation bar with links for CHINA NATIONAL GENE BANK, CNGB GLOBAL APPLICATION, SERVICES, and LOGIN/SIGNUP. The main header is "10KP Beta". Below the header is a search bar with the placeholder "Species, Sample name, Sample Id" and a magnifying glass icon. The main content area has several tabs: HOME (selected), STRATEGY, TIMELINE, SPECIES DATABASE, PRIORITY LIST, USER CENTER, and SAMPLE SUBMISSION. The "SAMPLE SUBMISSION" tab is currently active. It contains two main sections: "Online submission" and "Batch submission".

**10KP: 10,000 Plant Genomes Project**

The 10,000 plants (tenKP or 10KP) aims to sequence over 10,000 genomes representing every major clade of plants and eukaryotic microbes. This project would generate large-scale plant genome data within the next five years (2017-2022), addressing fundamental questions about plant evolution. Major supporters include Beijing Genomics Institute in Shenzhen (BGI-Shenzhen) and China National Gene Bank (CNGB). BGI corporate will support this project by developing new tools for de novo genome sequencing and assembly on MGISEQ platforms.

The announcement of the 10KP Project was published on July 27, 2017, 8:00 AM in Science. The establishment of this project is built on the success of the 1000 plants project, which was sampling phylogenetic diversity, not just crops and model organisms. The 10KP project would continue this strategy, and acquire new genome sequence information from the entire plant kingdom. For land plants, there are in total over 380,000 species, 26,700 genus from about 667 families, which are mostly from non-flowering plants (mosses, liverworts, hornworts, monilophytes/ferns, lycopodiophytes, gymnosperm, etc.) and flowering plants (magnoliids, basal angiosperms (ANA Grade), monocots/grasses, asterids, and rosid, etc.). For microbial eukaryotes, major attentions will focus on macro-/micro-algae and phototrophic/heterotrophic protists.

The 10KP Project will be a key part of the Earth BioGenome Project (EBP), an ambitious scheme to get at least rough sequence data from 1.5 million eukaryotic species. Currently the 10KP is open to receive plant samples from the world. The sample selection was based on a series of overlapping sub-projects with scientific objectives that could be addressed by the sequencing of multiple plant species. We are honored and looking forward to your participation!

**Online submission**

Please register your personal account through the CNGB Login System (automatic guided to registration when you click the new submission button below).

Provide all the information required for 10KP sample submission via a web-form including:

1. the basic information about the provider,
2. taxonomy of the plant species,
3. sample information (sample type & tissue type),
4. sample quality (concentration & OD value) and other information.

The administrator would review your submission and get back to you as soon as possible.

I have read 10kp's [Material Acquisition Agreement](#)

[+ New Submission](#)

**Batch submission**

For 10KP sample batch submission, please download and fill the submission template (10KP\_sample\_batch\_submission\_template.xls), and submit to the database.

\*10KP\_sample\_batch\_submission\_template.xls [Download](#)

\*Fill the submission template and submit to the website. Allowed file formats (.xlsx)

[Upload Samples](#)

Photo file name must correspond to the sample name. Please compress and upload the photo to the website. Single limit of 3 photos. Picture format

[Upload Photos](#)

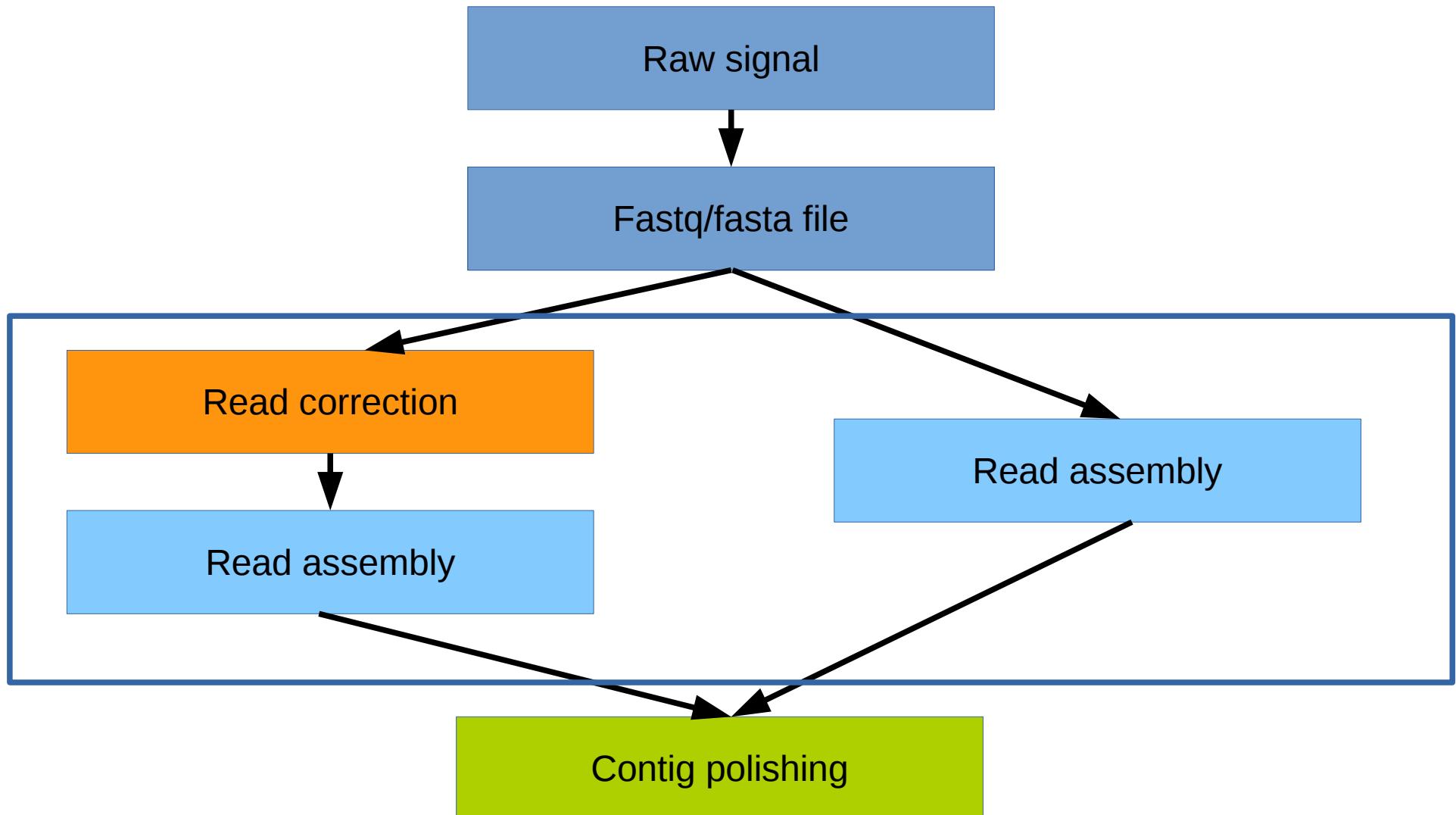
I have read 10kp's [Material Acquisition Agreement](#)

[Submit](#)

# 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

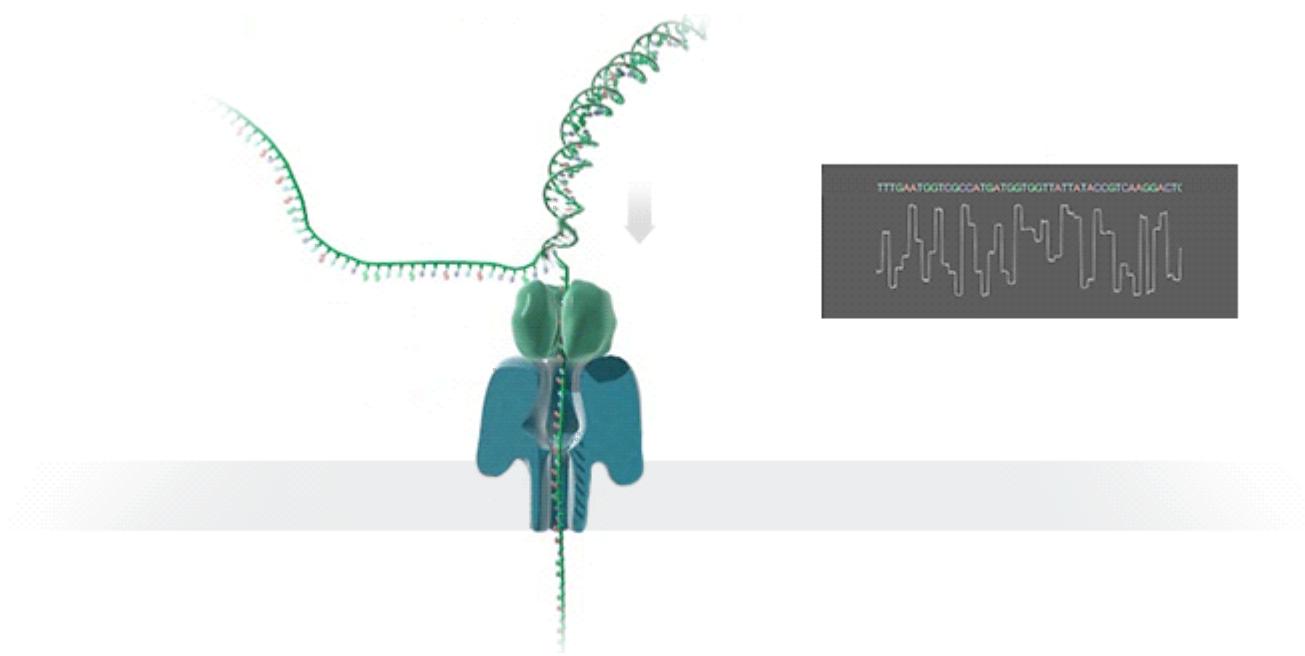


RSII



Sequel

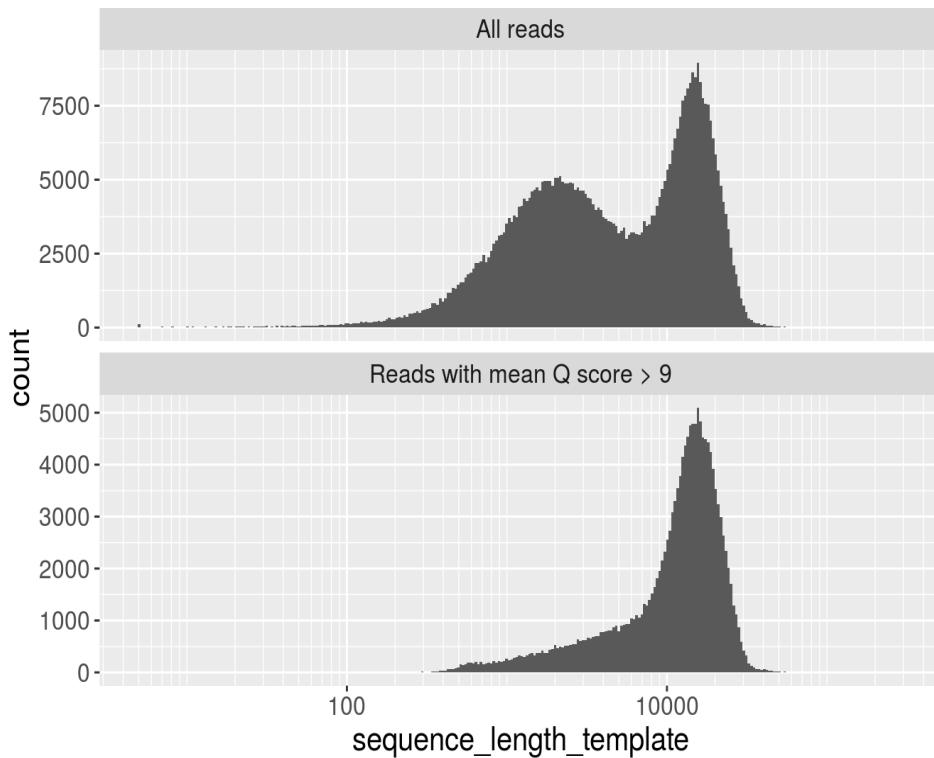
# ONT technology



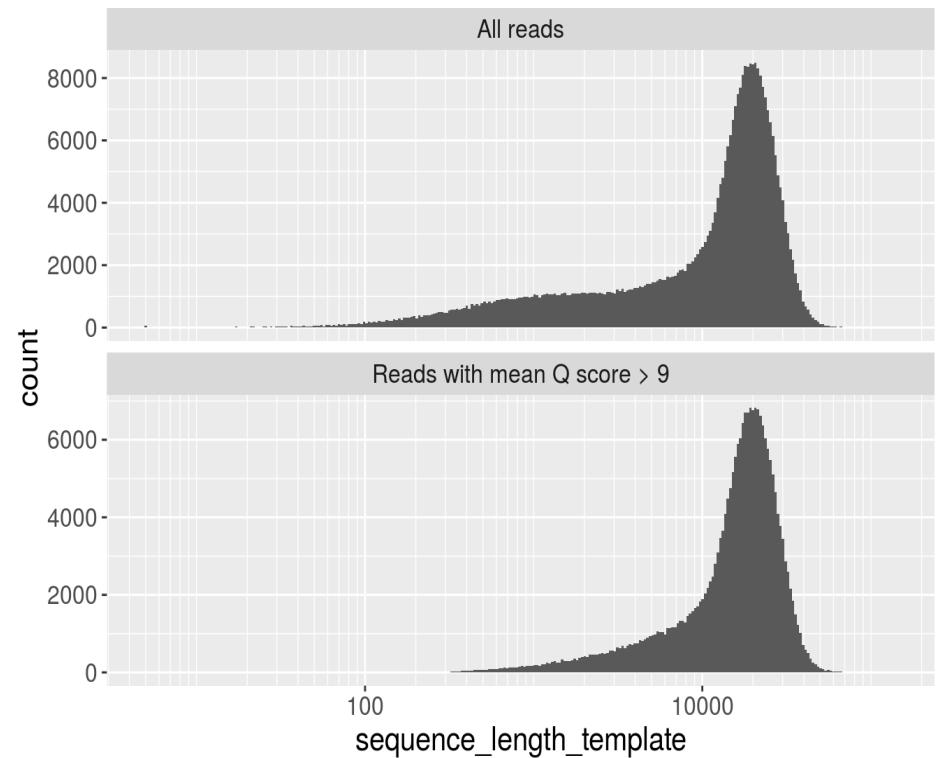
The signal is measured from five bases at a time.  
But timing is irregular.

# Read length and score filtering

Bad run



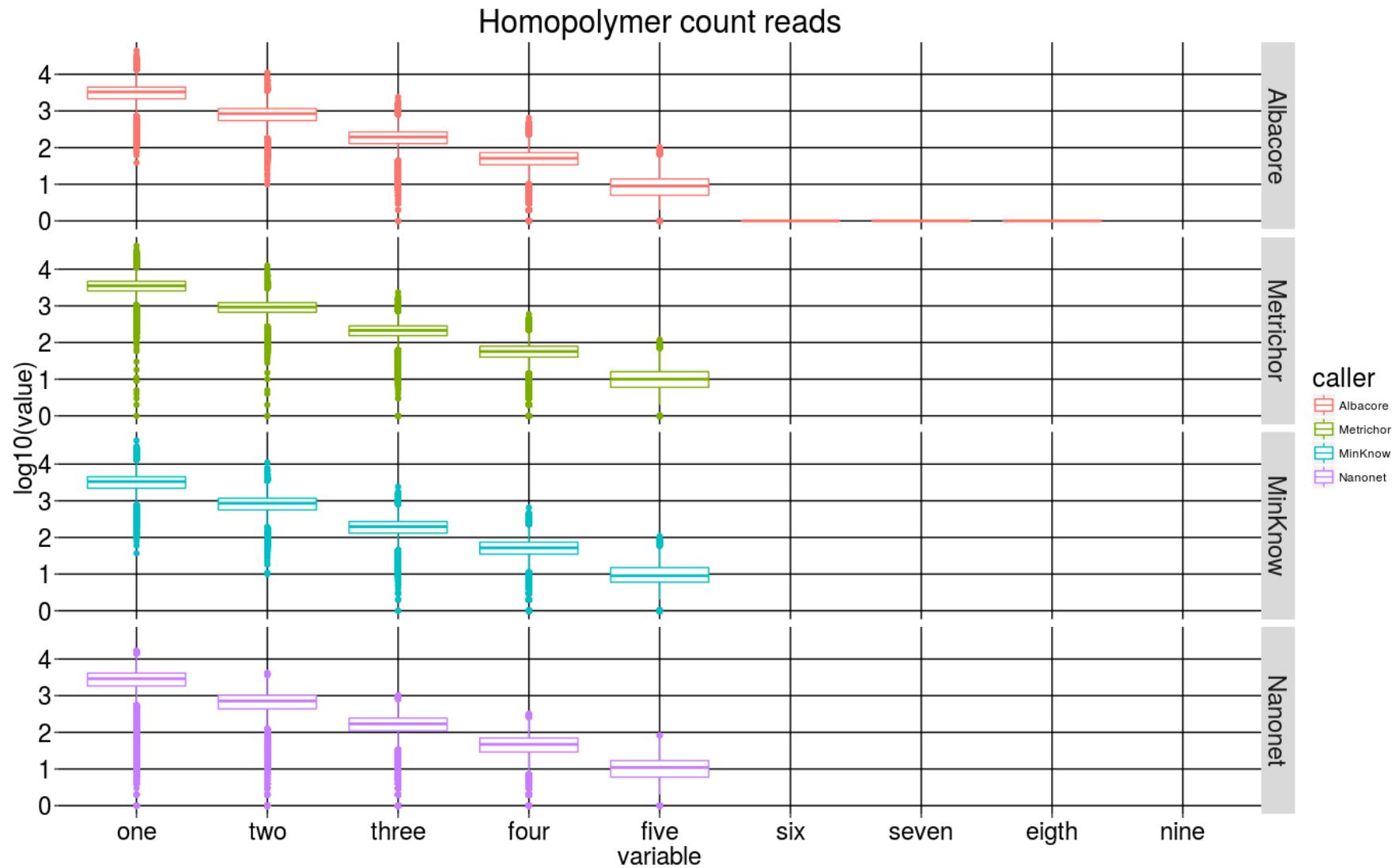
Good run



# ONT Read calling, cleaning and filtering

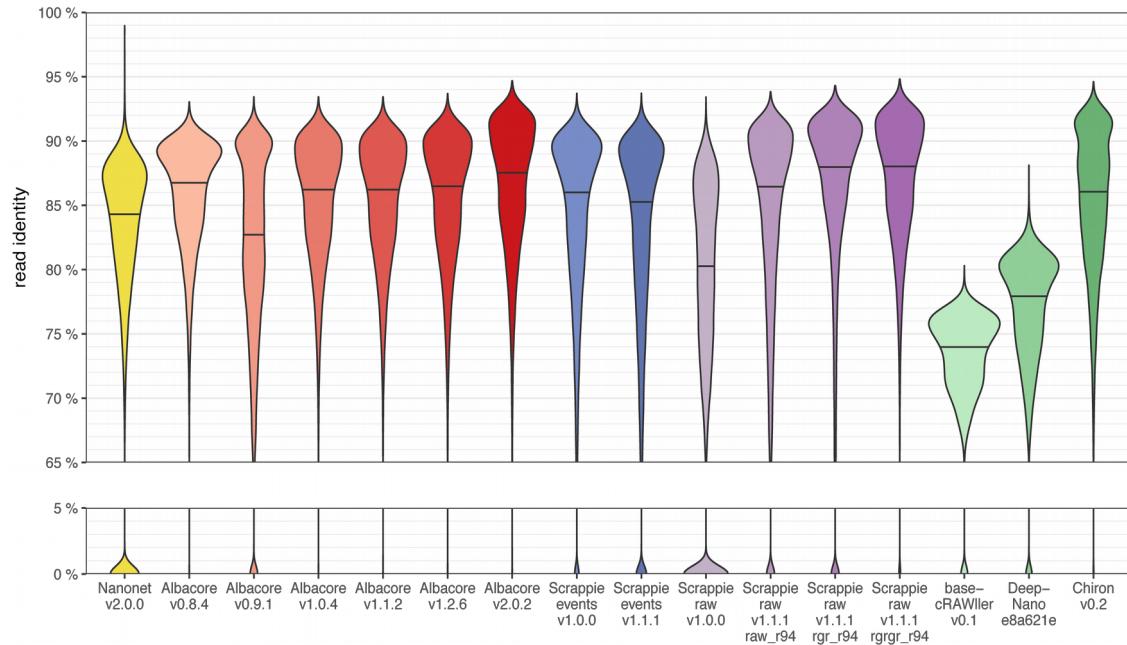
- Sequencer : raw fast5 files
- Step 1 : producing sequence information from fast5
  - **Albacore** : on the cluster or local server
- Step2 : extracting fastq from fast5
  - **Poretools** <https://github.com/arq5x/poretools>
- Step 3 : quality filtering using the sequencing\_summary.txt information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

# Homopolymers in reads

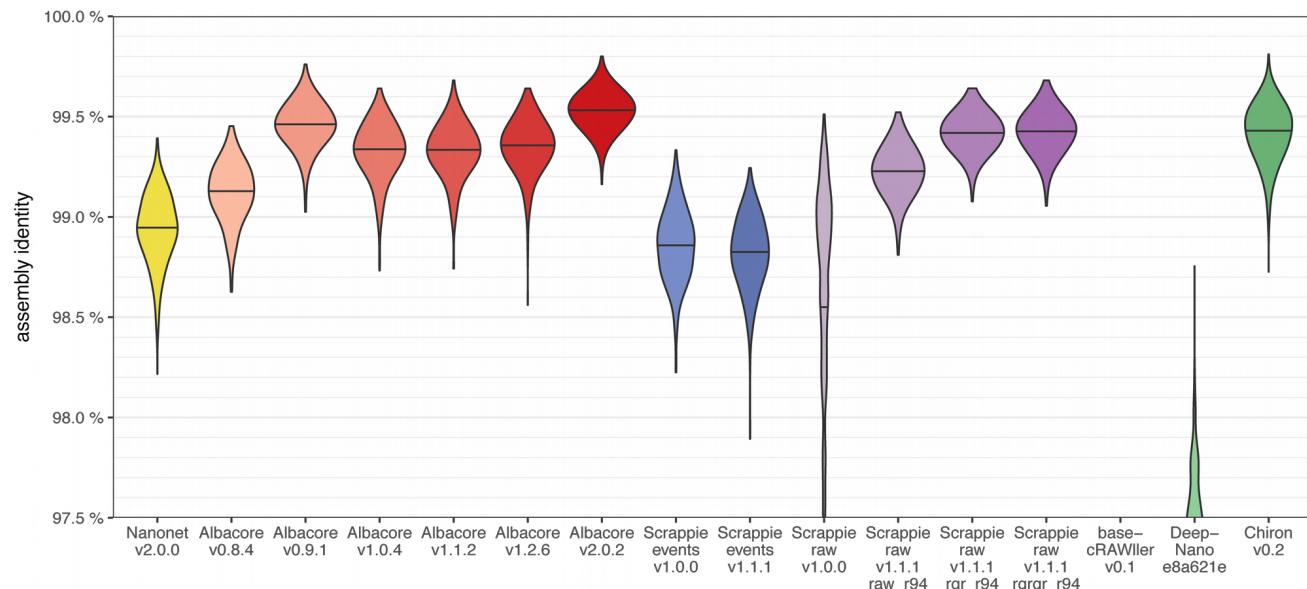


<https://github.com/rrwick/Basecalling-comparison>

Read identity



Assembly identity



# ONT Read analysis conclusions

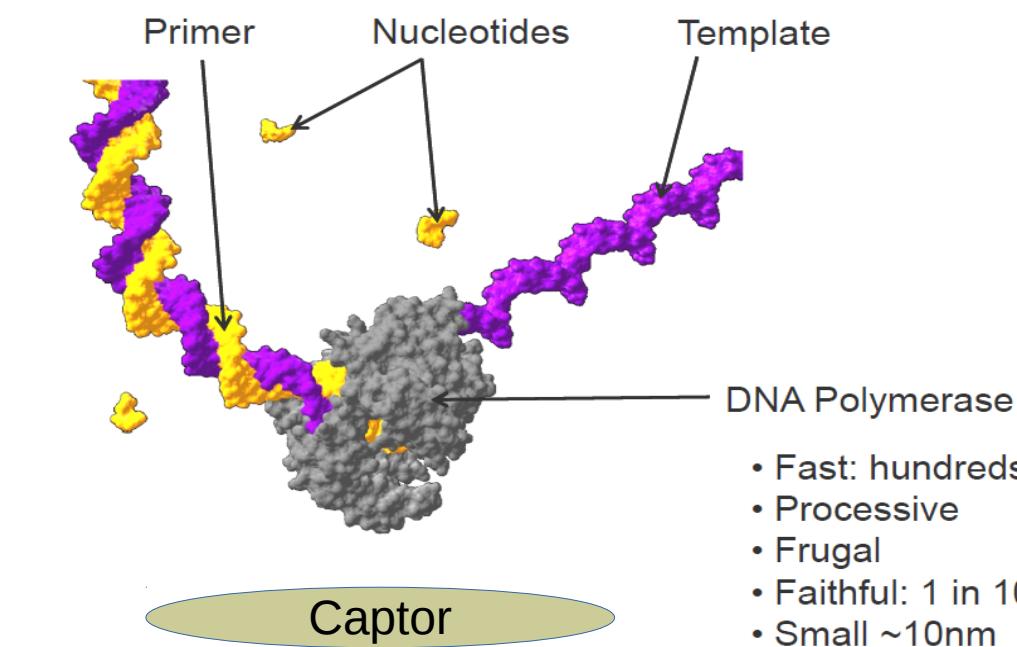
- Production up to 150Gb per cell for Promethion (15Gb for MinION and GridION).
- The error profile contains :
  - 6% of deletions
  - 4% of substitutions
  - 2% of insertions
- Follow the caller version updates to see if the next version are able to produce the long homopolymers
- Remove adapters with 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

# Circular consensus sequences

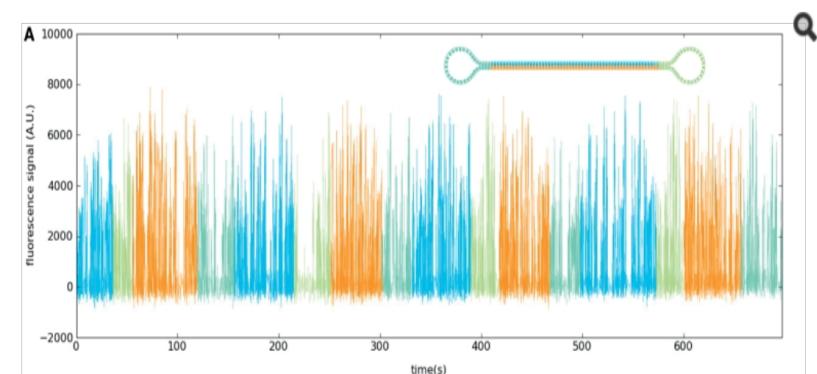
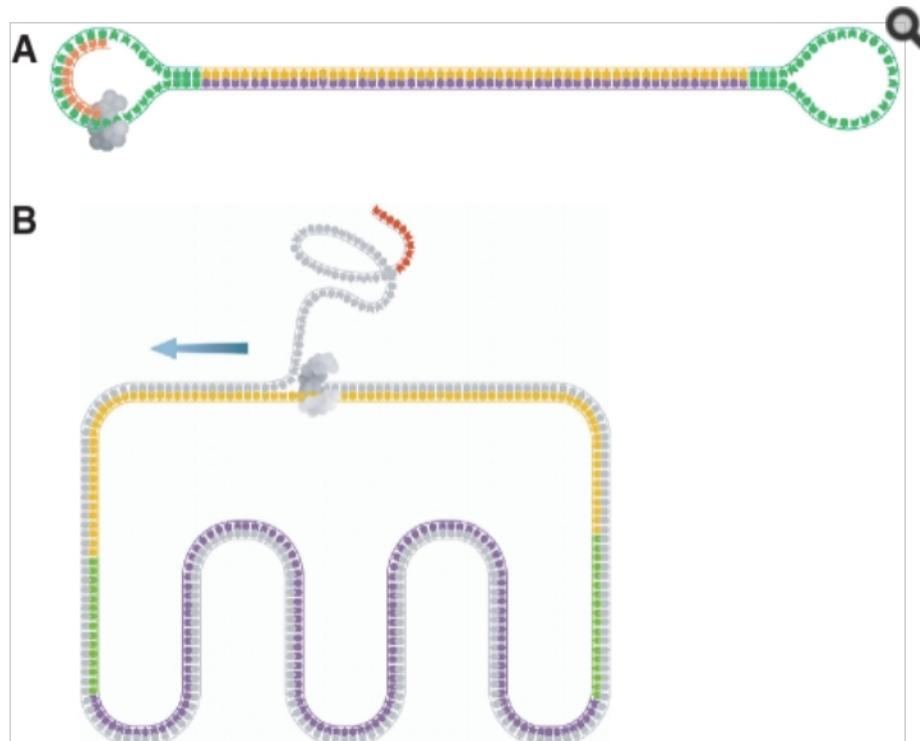
PMC full text: [Nucleic Acids Res. Aug 2010; 38\(15\): e159.](#)

Published online Jun 22, 2010. doi: [10.1093/nar/gkq543](https://doi.org/10.1093/nar/gkq543)

[Copyright/License ▶](#)

[Request permission to reuse](#)

**Figure 1.**



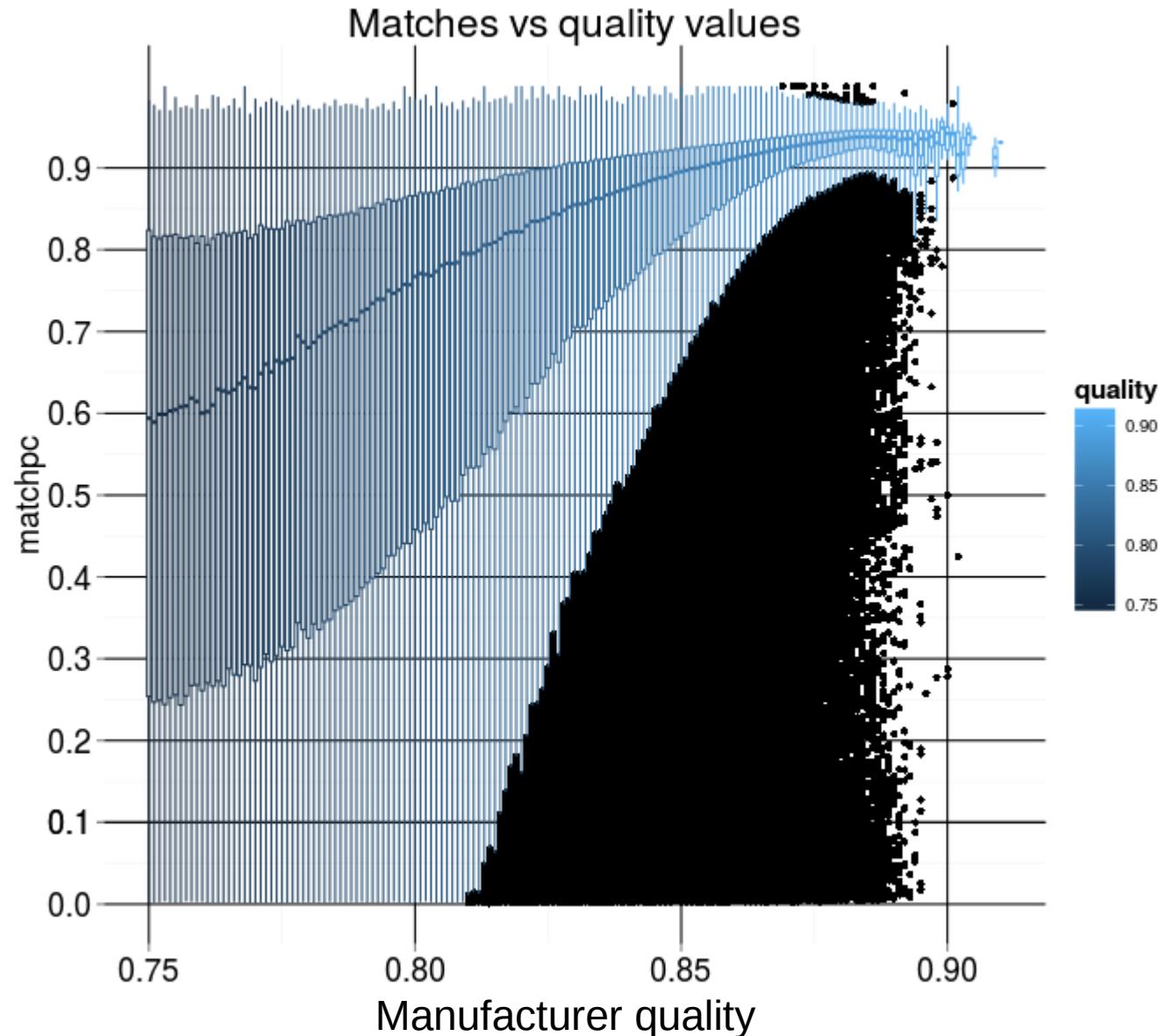
<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2926623/>

# 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)

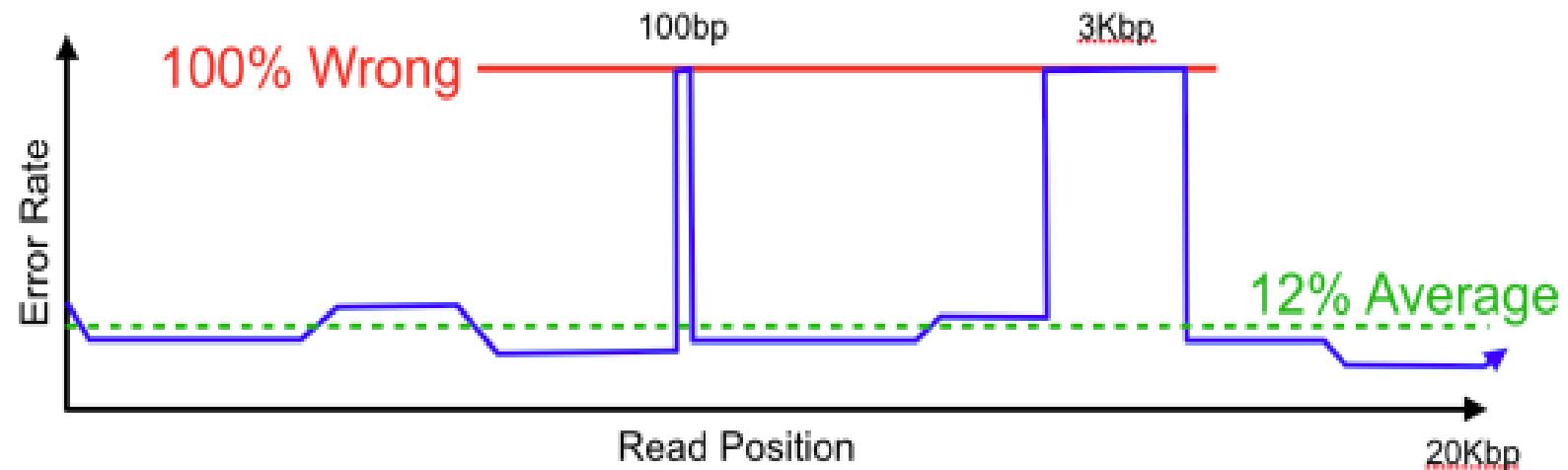
```
@m150619_093250_42174_c100795682550000001823166309091510_s1_p0/109/0_4936 RQ=0.879
@m150619_093250_42174_c100795682550000001823166309091510_s1_p0/109/4981_9942 RQ=0.879
@m150619_093250_42174_c100795682550000001823166309091510_s1_p0/109/9988_10378 RQ=0.879
@m150619_093250_42174_c100795682550000001823166309091510_s1_p0/157/0_7588 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

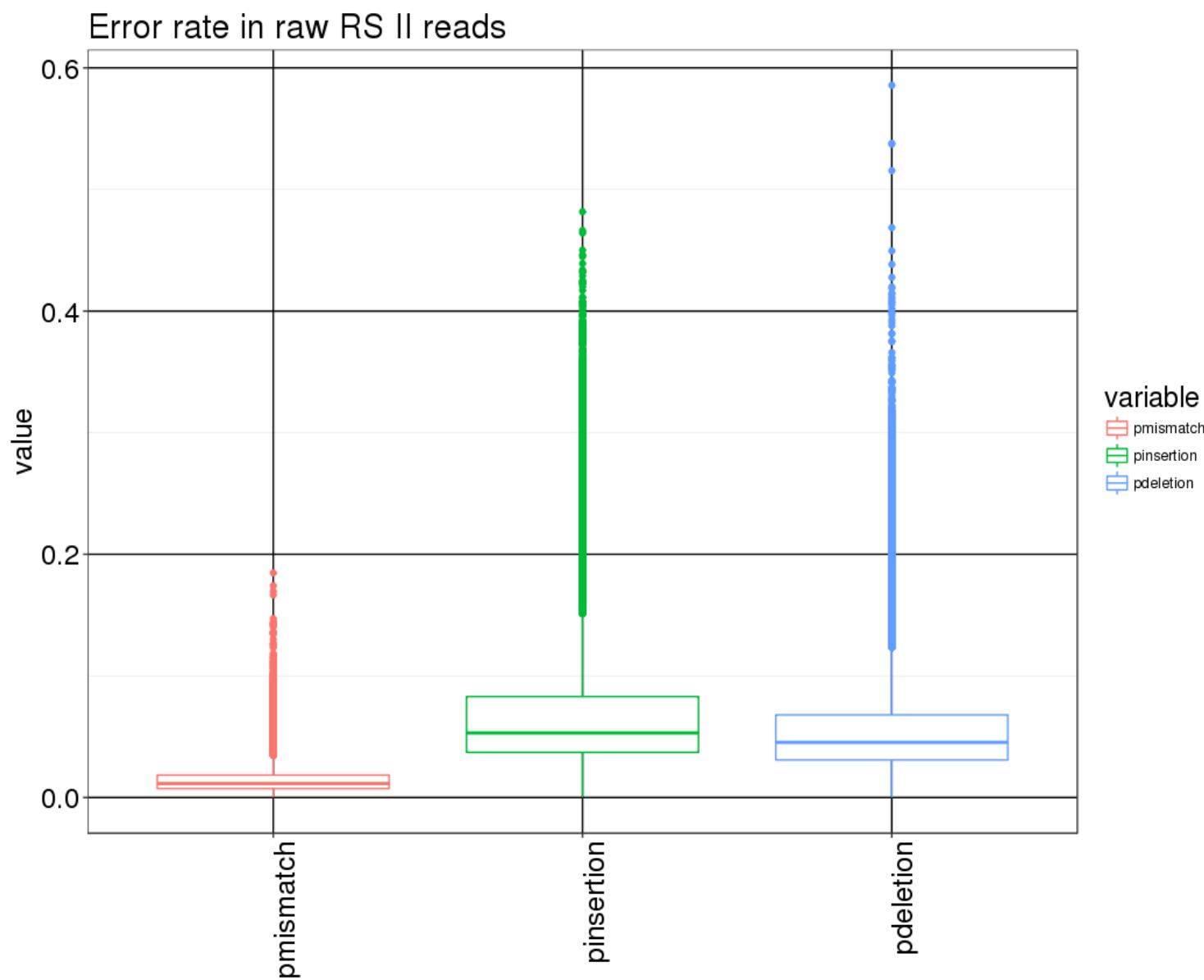


# Polymerase errors

<https://dazzlerblog.wordpress.com/>



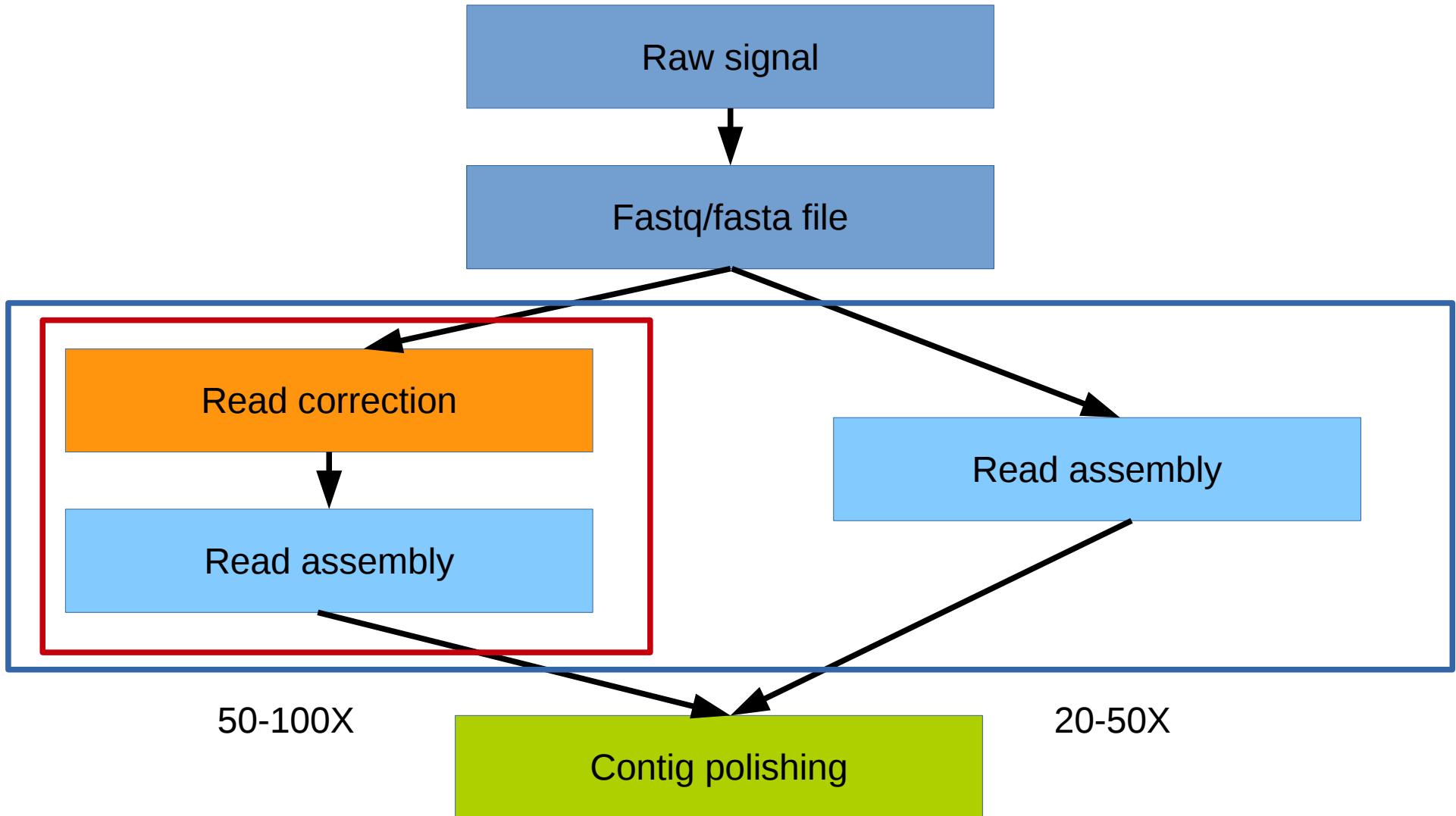
# PacBio RSII error profile



# 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)

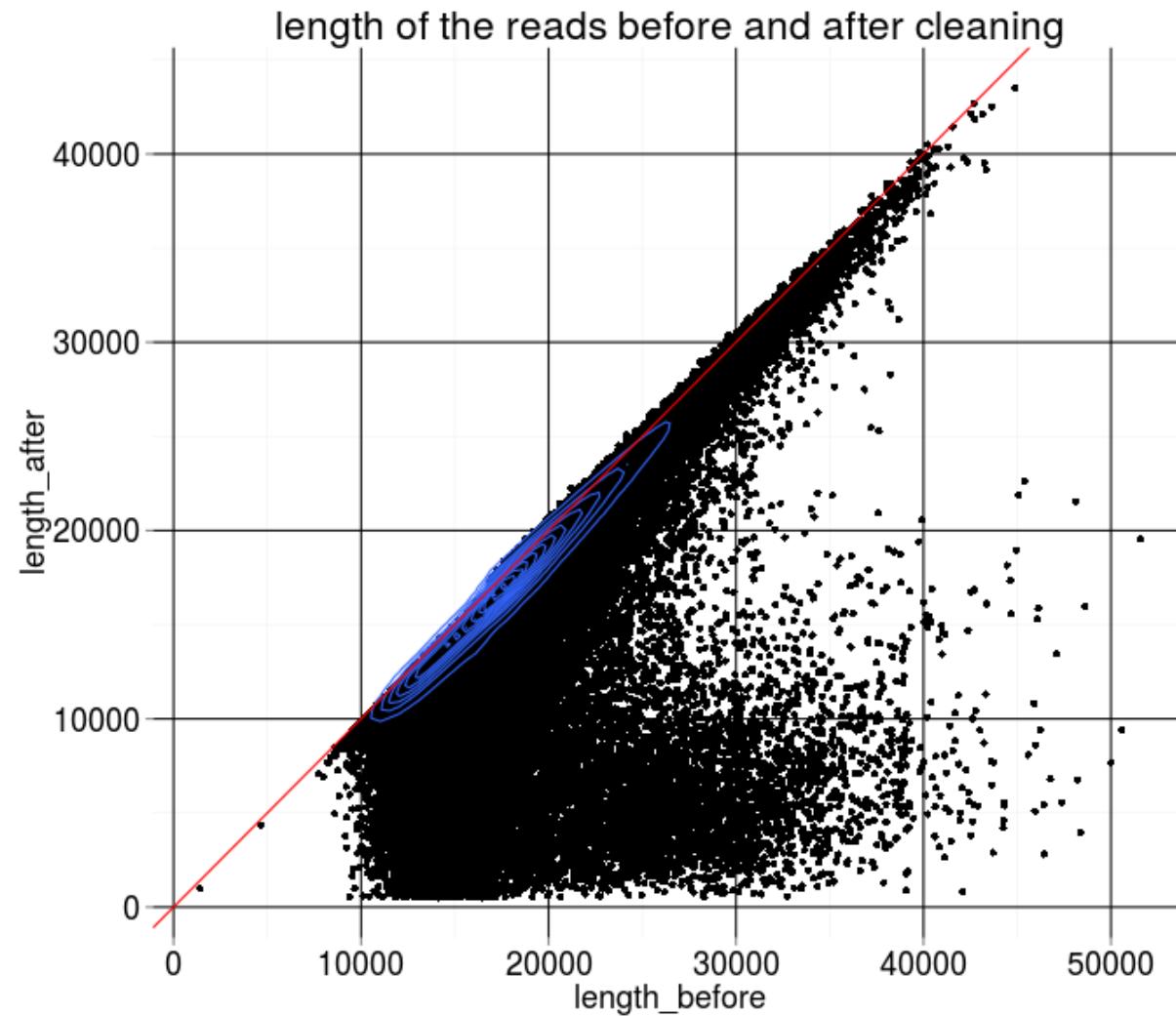
# Assemblers with read correction



# 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

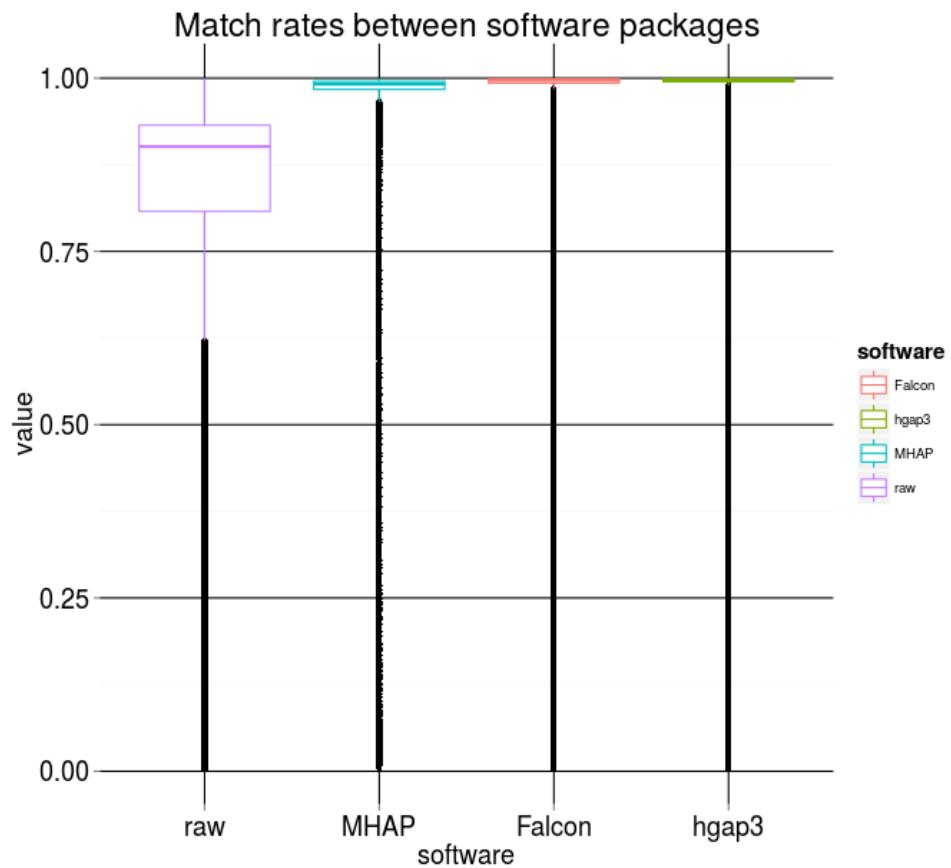
# Corrected versus raw lengths



# Assemblers with read correction

	HGAP	CANU	Falcon
Read correction <b>aligner</b>	Read splitting for correction. <b>blasr</b>	<b>MHAP</b>	1 read per film selection <b>daligner</b>
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.
Falcon	0.004315	0.992900	0.996900	0.969200	0.998600	1.000000

**hgap3**

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
hgap3	0.003802	0.994700	0.997100	0.969100	0.998300	1.000000

**MHAP**

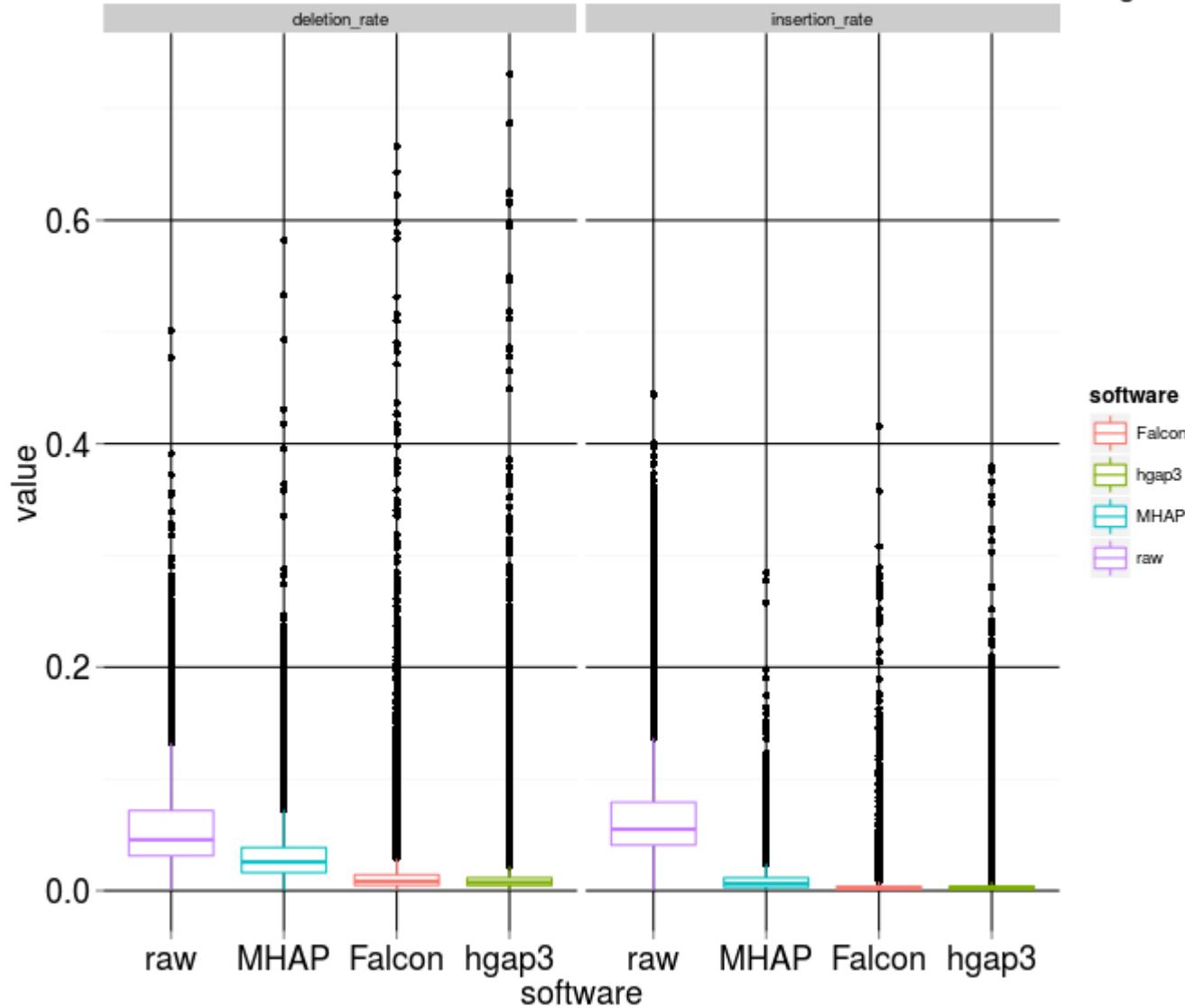
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
MHAP	0.005378	0.983600	0.991300	0.966600	0.995600	1.000000

**raw**

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

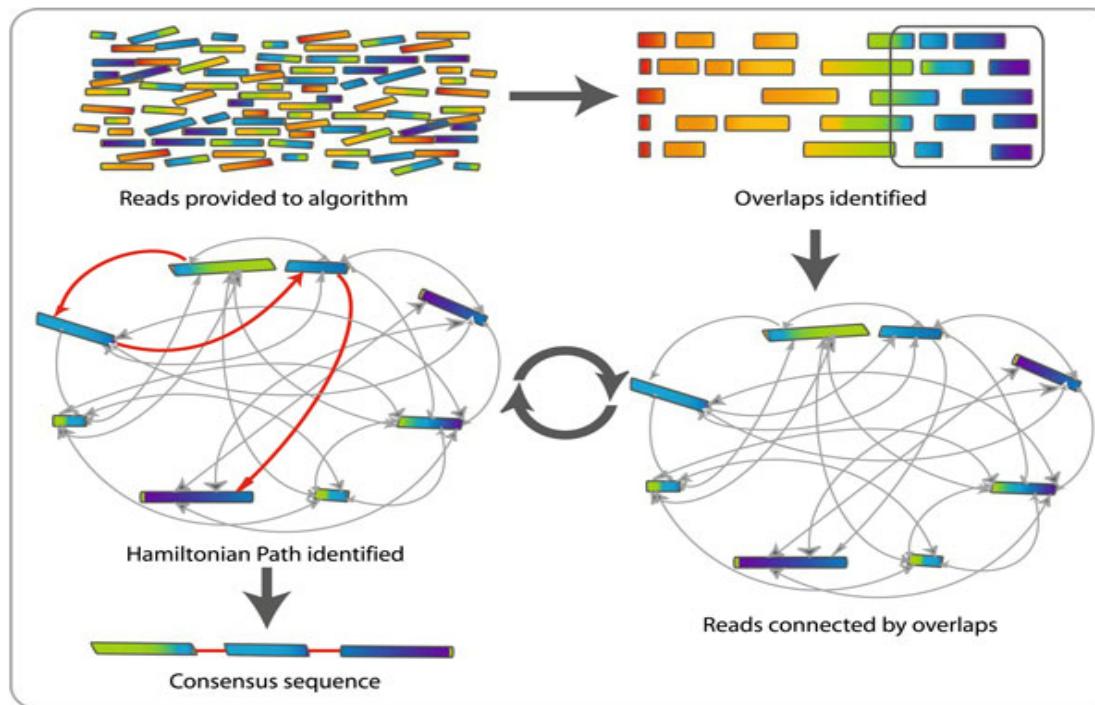
# Insertion and deletion rates

Evolution of the insertion and deletions of the reads before and after cleaning



# Assembling corrected reads

- OLC (Overlap Layout Consensus) assembler



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

- Celera assembler used by hgape, Canu, Falcon

# Contig assembly metrics

Species Assembler	Genome size	Coverage	Total contig size	Contig count	N50	L50
Caenorhabd itis elegans <b>Falcon PB</b>	100 Mb	47X	101,72 Mb	122	2,022,653	17
Aeschynom ene evenia <b>Hgap PB</b>	400 Mb	120X	374.45 Mb	5,711	648,407	87
Ictalurus punctatus <b>Falcon PB</b>	900 Mb	120X	826,04 Mb	1 ,554	4,431,159	50
Oryzias javanicus <b>Falcon PB</b>	900 Mb	50X	865,44Mb	1,286	3,821 ;811	59
Arabidopsis thaliana <b>Canu ONT</b>	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
<b>Caenorhabditis elegans</b> <b>Miniasm PB</b>	100 Mb	47X	106,81 Mb	129	2,214,883	17
<b>Oryzias javanicus</b> <b>Miniasm PB</b>	900 Mb	50X	816,95 Mb	2,638	685,722	323

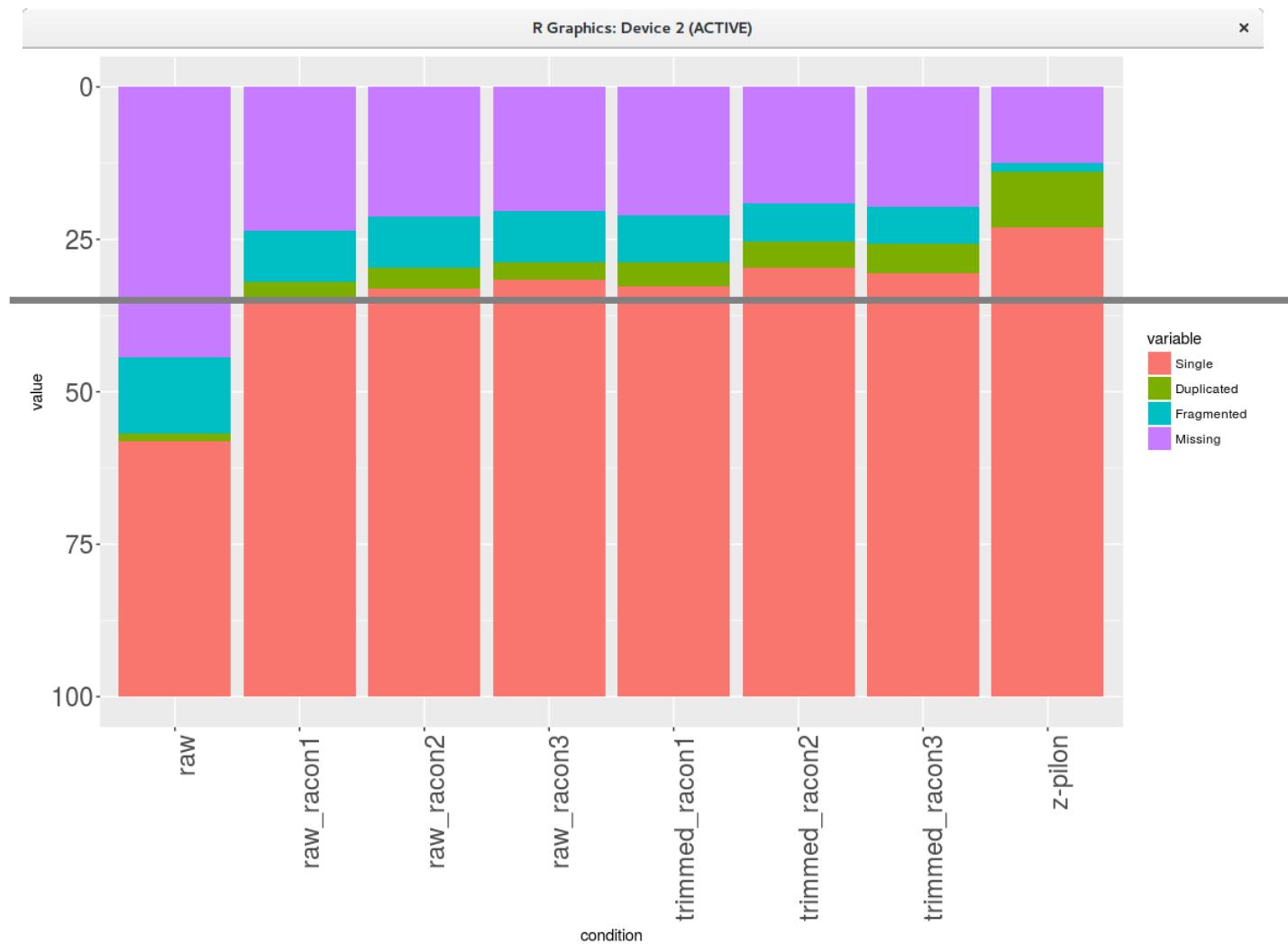
# Assembly polishing

- 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 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

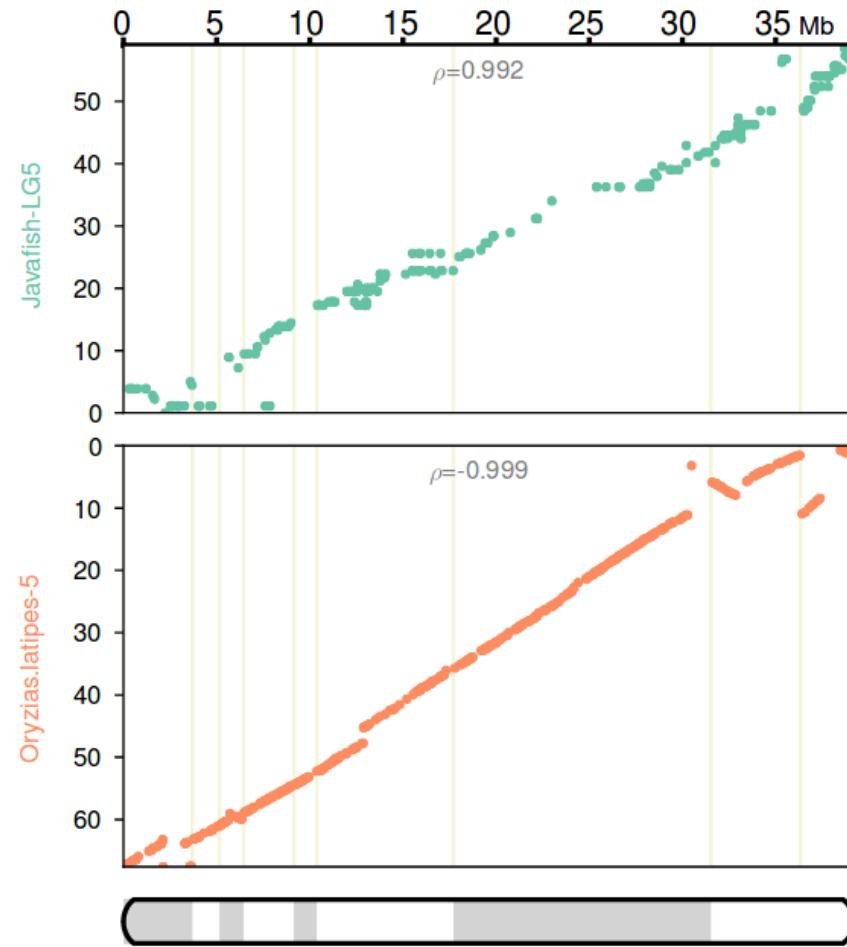
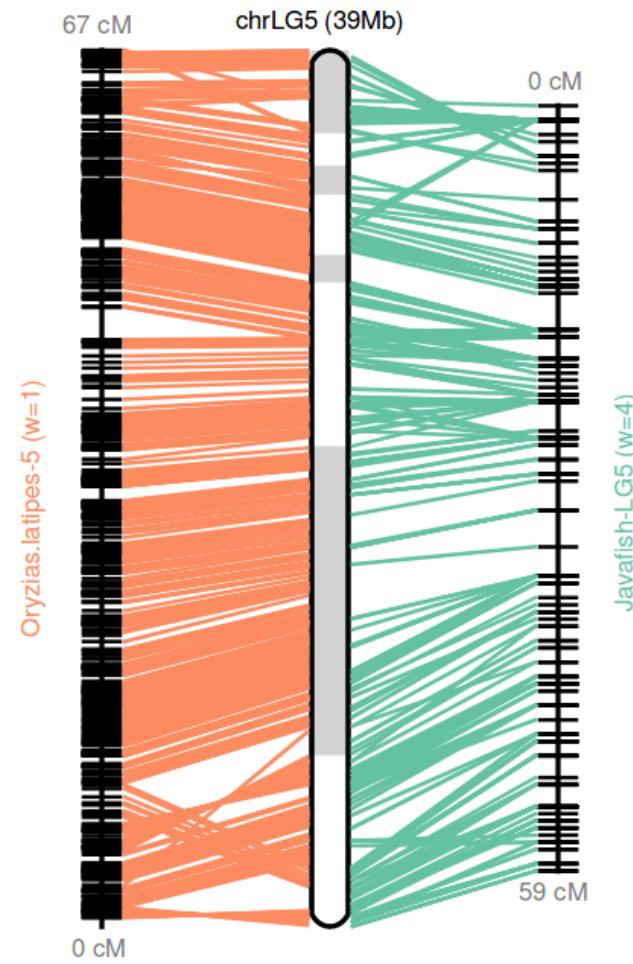
<https://busco.ezlab.org/>



# From contigs to chromosomes

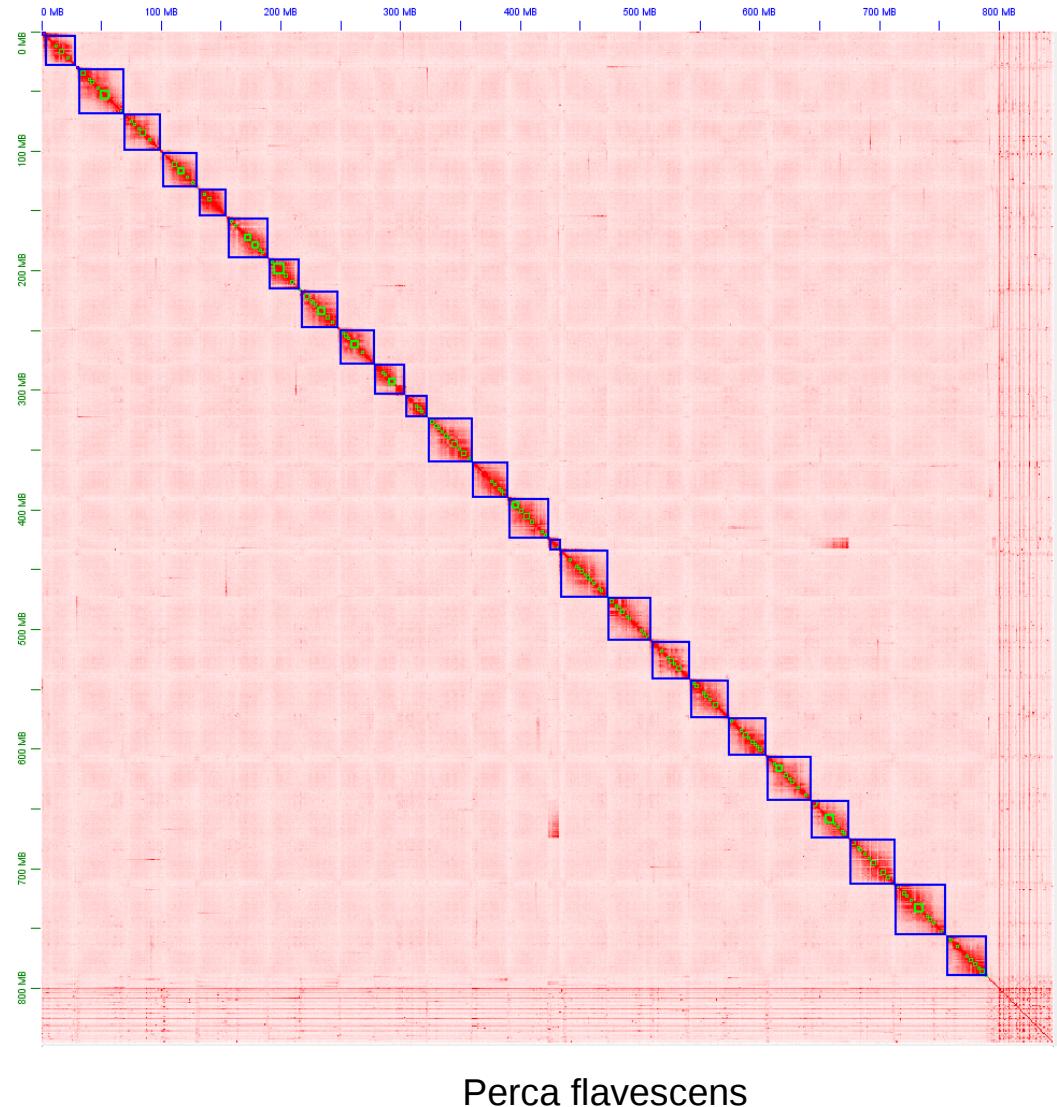
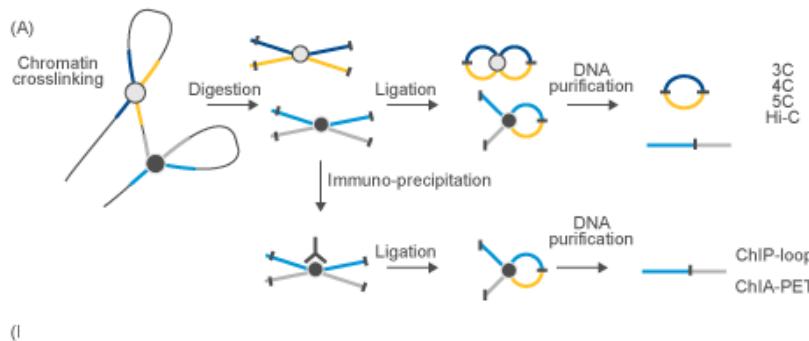
- **Optical mapping** : fluorescent marking of restriction sites of very long DNA molecules (up to Mb) to extract signature used to bridge contigs having these signatures.
- **10x chromium** : shallow tagged sequencing of very long DNA fragments with Illumina machines. Read alignments enable scaffolding.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

# Building chromosomes : genetic map



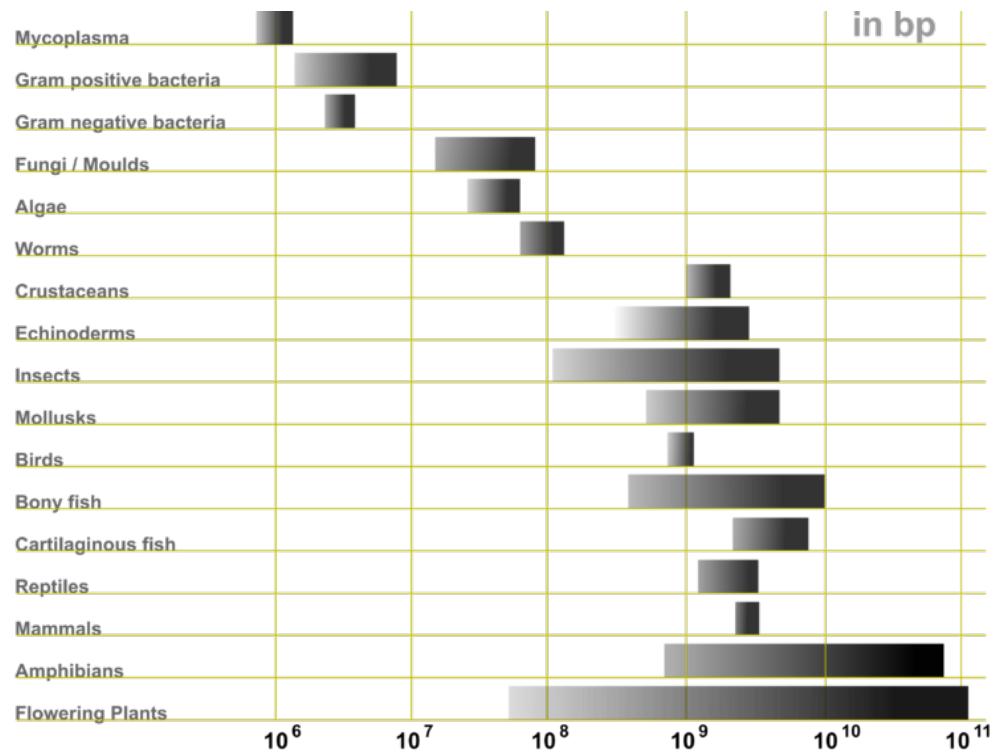
# Building chromosomes : Hi-C

## Hi-C protocol



# What makes genomes difficult to assemble ?

- Genome size
- Number of chromosomes
- Repeat content
- Repeat size (structure)
- Heterozygosity
- Ploidy
- DNA conformation
- Contamination



[https://en.wikipedia.org/wiki/Genome\\_size](https://en.wikipedia.org/wiki/Genome_size)

# Sturgeon karyotype

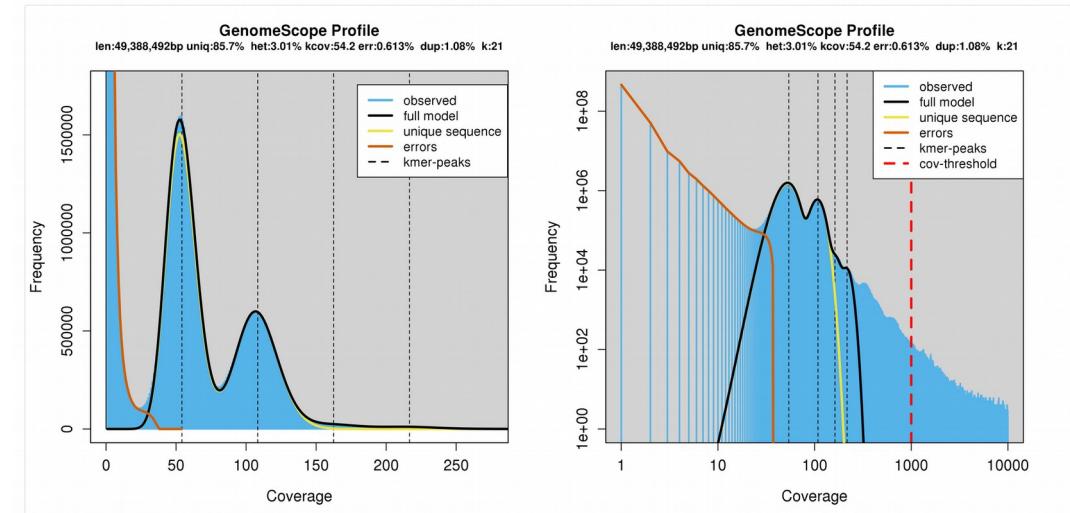


[http://www.unife.it/dipartimento/biologia-evoluzione/progetti/geneweb/immagini/babur73\\_g.jpg](http://www.unife.it/dipartimento/biologia-evoluzione/progetti/geneweb/immagini/babur73_g.jpg)

$2n \sim 209 \dots 249$

# Difficult to assemble genomes 1

Polyplloid genomes  
Ganoderma fungi  
dikaryotic stage



	MiSeq	PacBio
software	DiscoverDeNovo	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

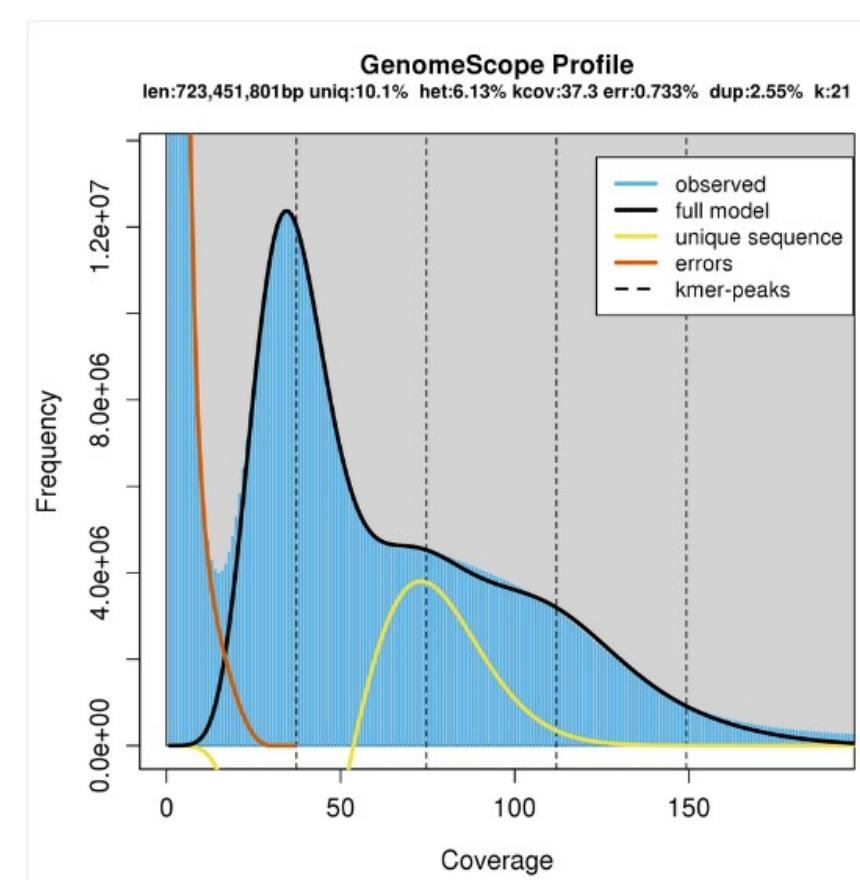
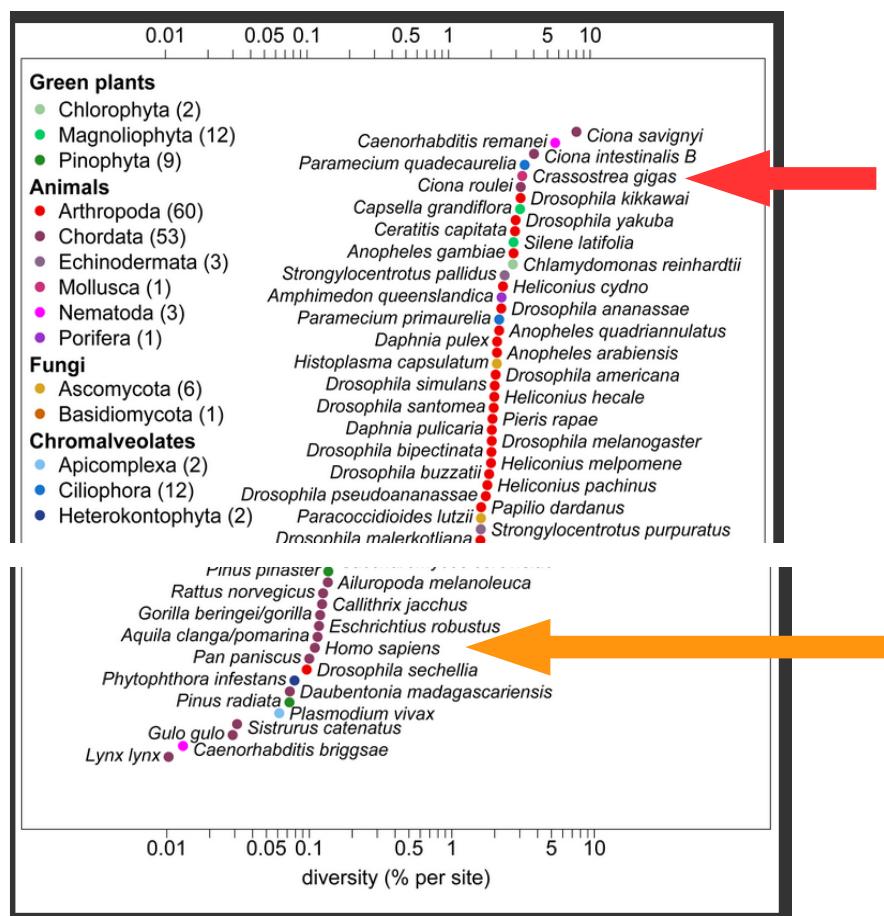
GenomeScope version 1.0  
k = 21

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	<b>1,068,631</b>	<b>889,398</b>
L50 contig count	55	17	18

# Difficult to assemble genomes 2

## Heterozygosity: oyster



# 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)



Jean-François Arrighi (IRD)

Sylvie Quiniou (USDA)

Roberto Lleras (PacBio)

Yann Guiguen (INRA)



Alain Vignal (INRA)



Jérôme Gouzy (INRA)



Cécile Donnadieu (INRA)

Alain Roulet (INRA)

Get-Plage team members