


Polishing of genome assemblies

EBAll Assembly & Annotation - Roscoff Juin 2024

Jean-Marc Aury

Laboratoire de Bio-informatique pour la Génomique et la Biodiversité

 jmaury@genoscope.cns.fr

 [@J_M_Aury](https://twitter.com/J_M_Aury)

What are you going to learn?

- What is polishing.
- How to spot a potential problem with your assembly consensus.
- How polishing tools work.
- Which are the most common polishing tools.
- How to polish a genome assembly.

What is polishing ?

Polishing is an important step in genome assembly that involves inspecting the consensus of a given assembly to detect local errors.

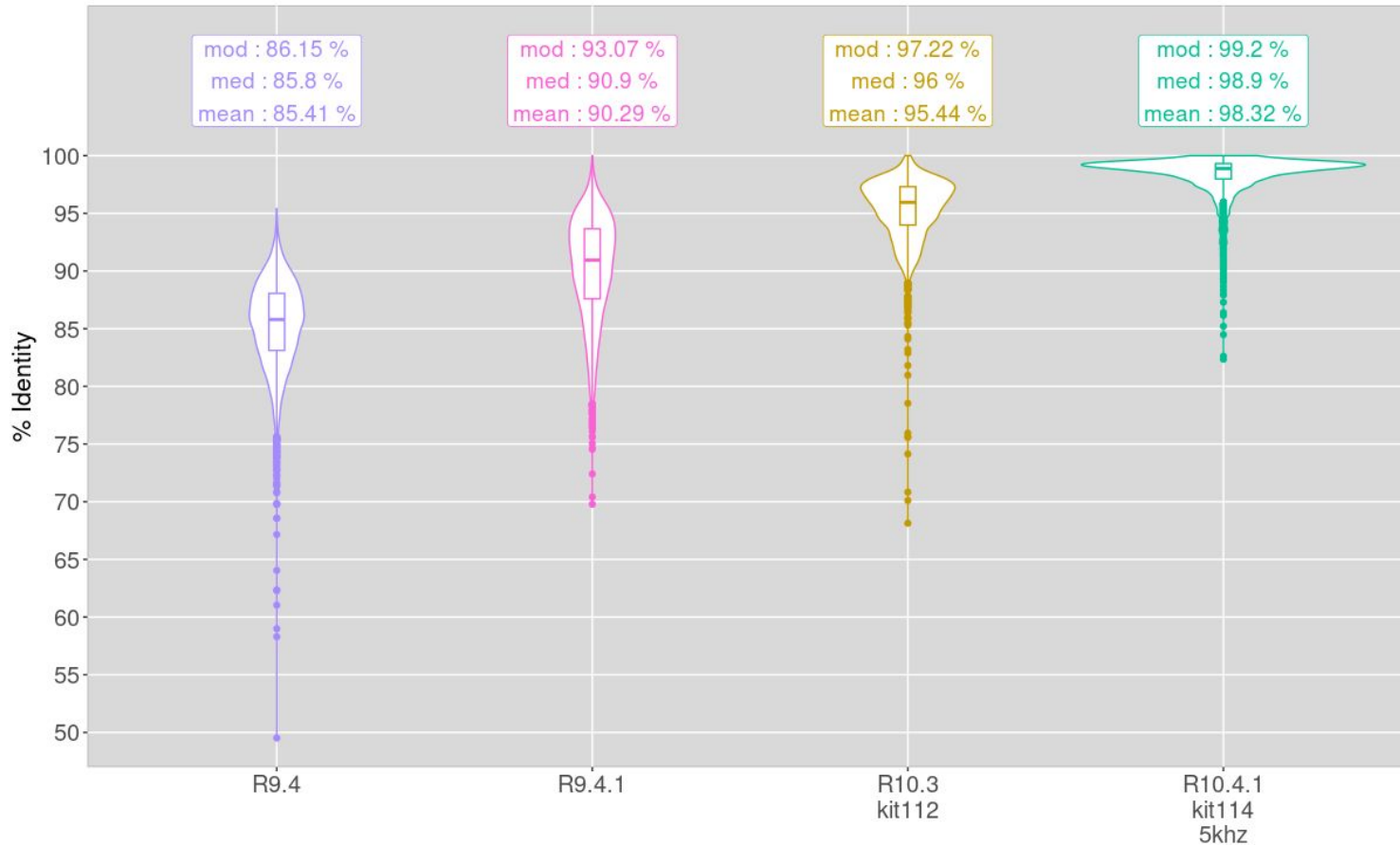
The “polishing” step generally requires high-quality reads (Illumina, MGI or PACBIO HiFi) and a genome assembly.

Why do we need to polish our assemblies ?

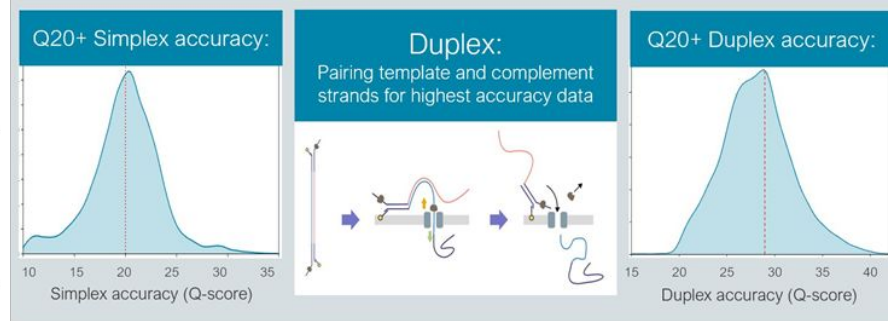
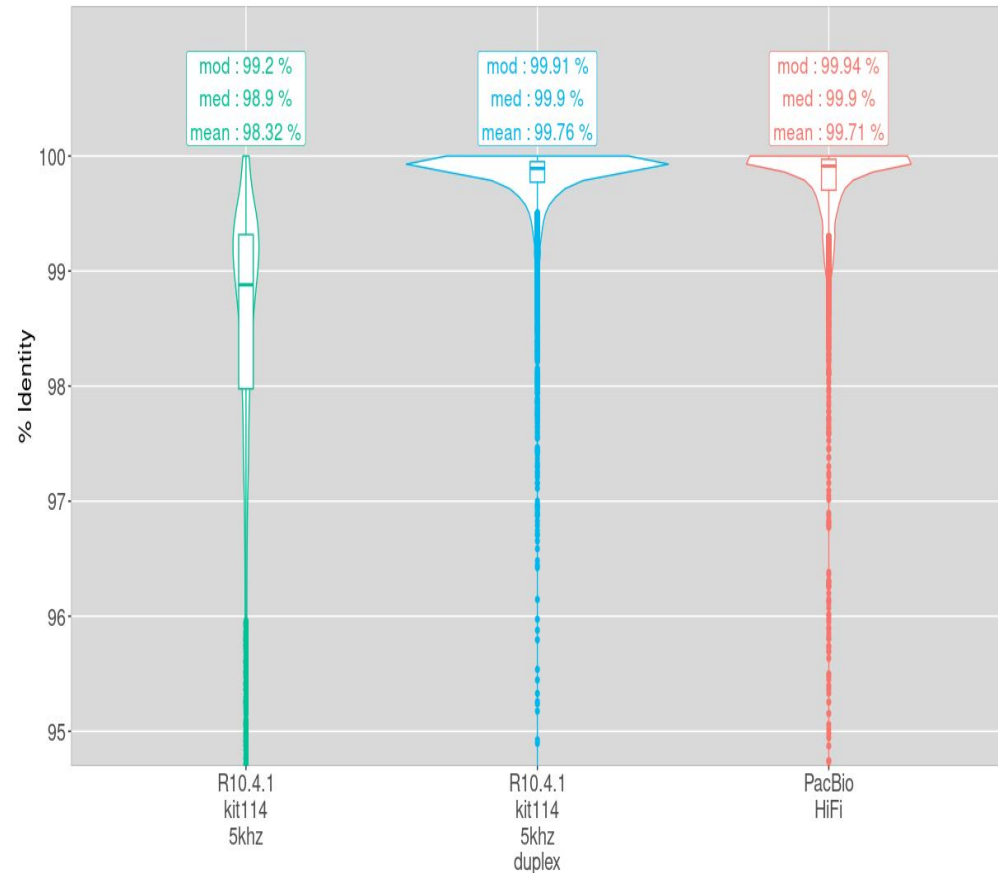
Due to sequencing error rate, the consensus of a given genome assembly might contains errors : mismatches, insertion or deletion

Insertion or deletion may affect the frame of coding sequences and result in incomplete gene prediction. This problem can be detected with BUSCO.

A fast evolving technology



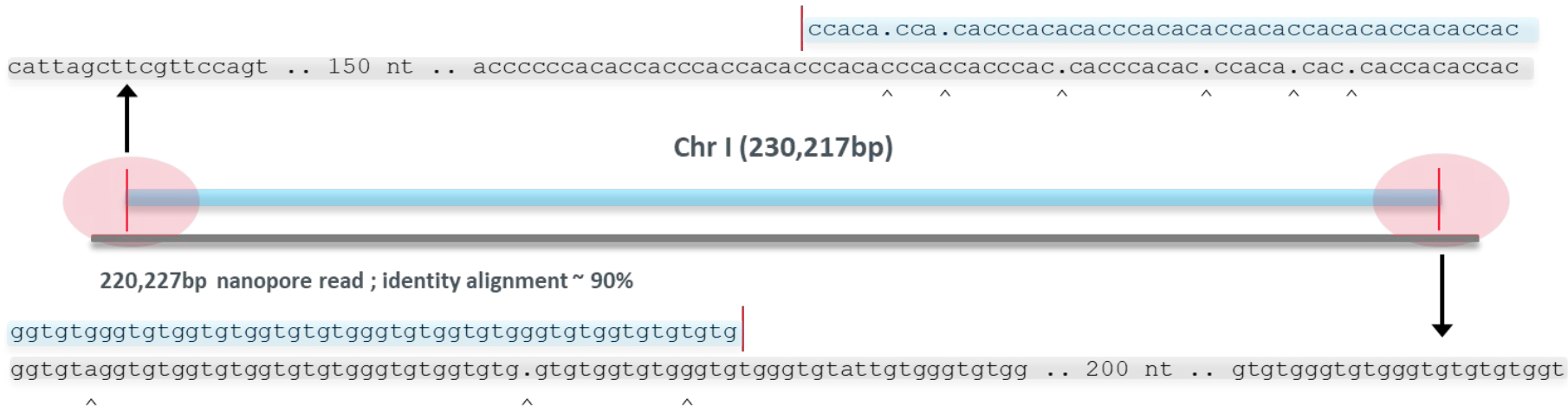
A fast evolving technology





- Sequencing kit114
- R10 Flowcells
- Guppy 6

A fast evolving technology

Chromosomes can be captured entirely, the example read span yeast chromosome 1 from telomere to telomere

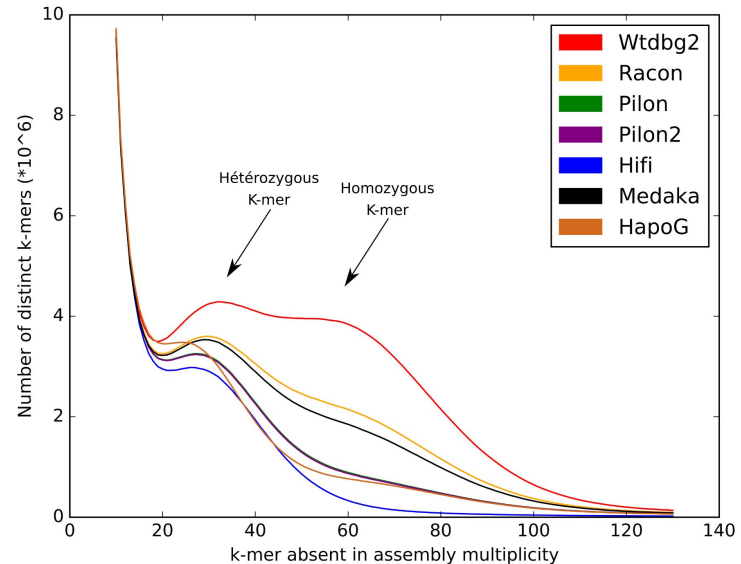
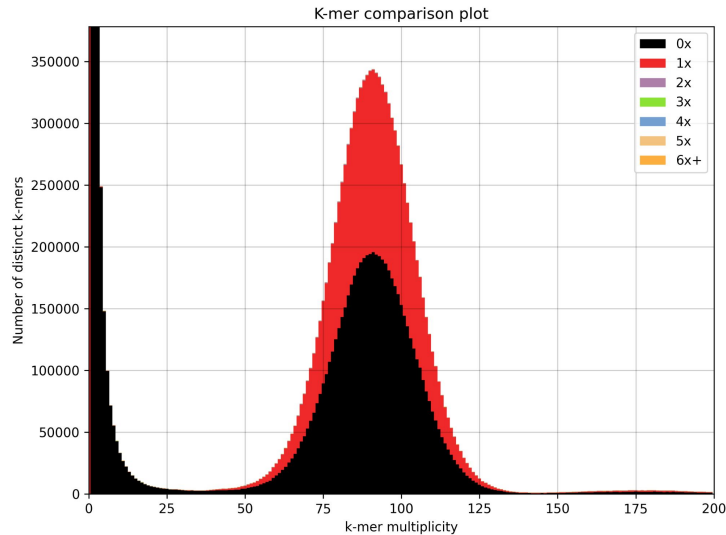


Read length from Nanopore and community

	
"Whale scale" 1 kg = 1 kbase	Int: 4.2 Mb human data Ext: 2.4 Mb human data

How to spot a potential problem with your assembly consensus.

- Each kmer of your readset should also be found in your genome assembly
=> Generate a KAT plot



How to spot a potential problem with your assembly consensus.

- Errors in your consensus can affect gene prediction
=> Launch BUSCO and look at the “Complete”, “Fragmented” and “Missing” scores

```
# BUSCO version is: 5.2.2
# The lineage dataset is: eukaryota_odb10 (Creation date: 2020-09-10, number of
genomes: 70, number of BUSCOs: 255)
# Summarized benchmarking in BUSCO notation for file
/env/export/bigtmp2/jmaury/ebaii/nanopore_assembly_flye/Assembly/Flye/nanopore.fasta
# BUSCO was run in mode: genome
# Gene predictor used: metaeuk
```

```
***** Results: *****
C:57.3%[S:57.3%,D:0.0%],F:12.2%,M:30.5%,n:255
146      Complete BUSCOs (C)
146      Complete and single-copy BUSCOs (S)
0        Complete and duplicated BUSCOs (D)
31       Fragmented BUSCOs (F)
78       Missing BUSCOs (M)
255      Total BUSCO groups searched
```


How to spot a potential problem with your assembly consensus.

- Quality score calculated by Merqury (using short-reads) will be low
=> Launch Merqury and look at the Quality Value

$Q = -10 \log_{10} P$ (Q= Quality value and P= basecalling error probability)

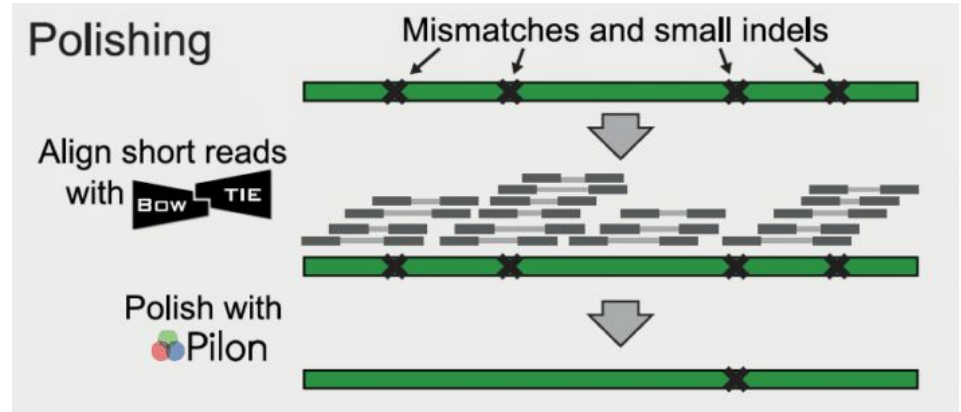
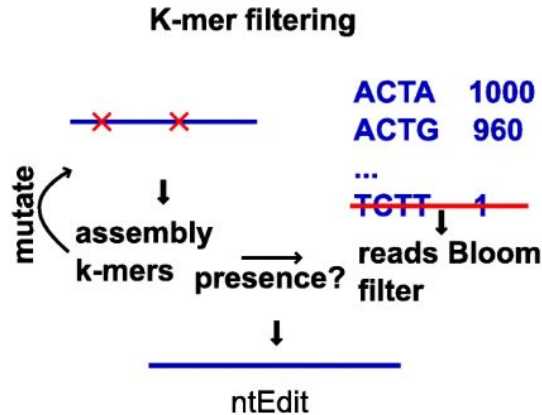
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10 Kb	99.99%
50	1 in 100 Kb	99.999%

```
[jmaury@inticns] ## cat flye/Merqury/merqury.qv
```

```
nanopore          4338717 9512352 17.1099 0.0194539
```

How polishing tools work.

- Two different strategies:
 - kmer-based approach : faster, but less accurate
 - alignment-based approach : slower, but more accurate

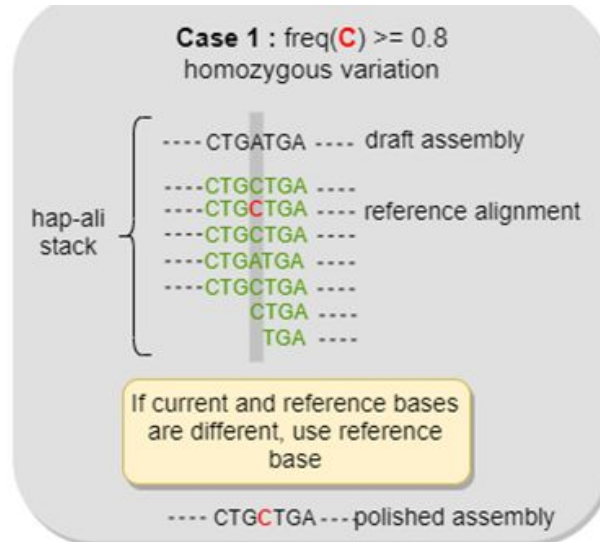


Amarasinghe, S.L., Su, S., Dong, X. et al. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol* 21, 30 (2020). <https://doi.org/10.1186/s13059-020-1935-5>

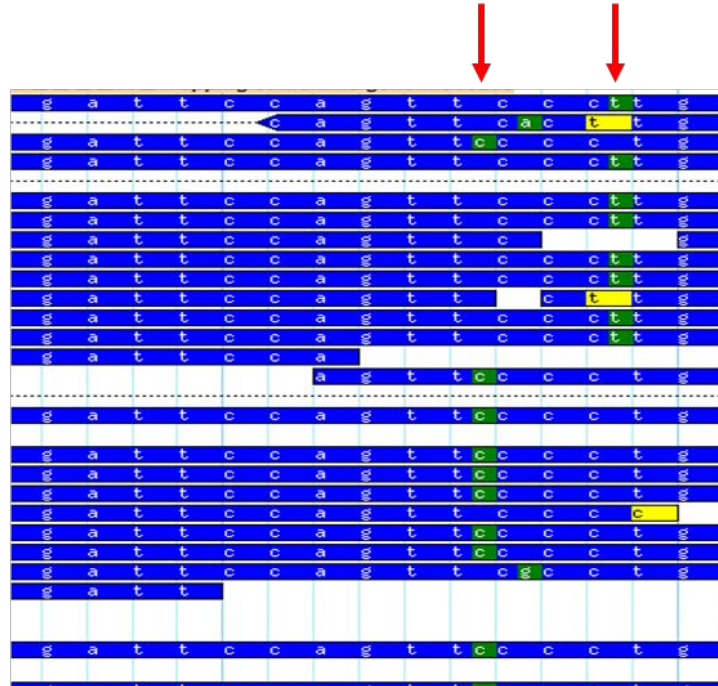
<https://thesequencingcenter.com/knowledge-base/complete-genome-assembly/>

How polishing tools work.

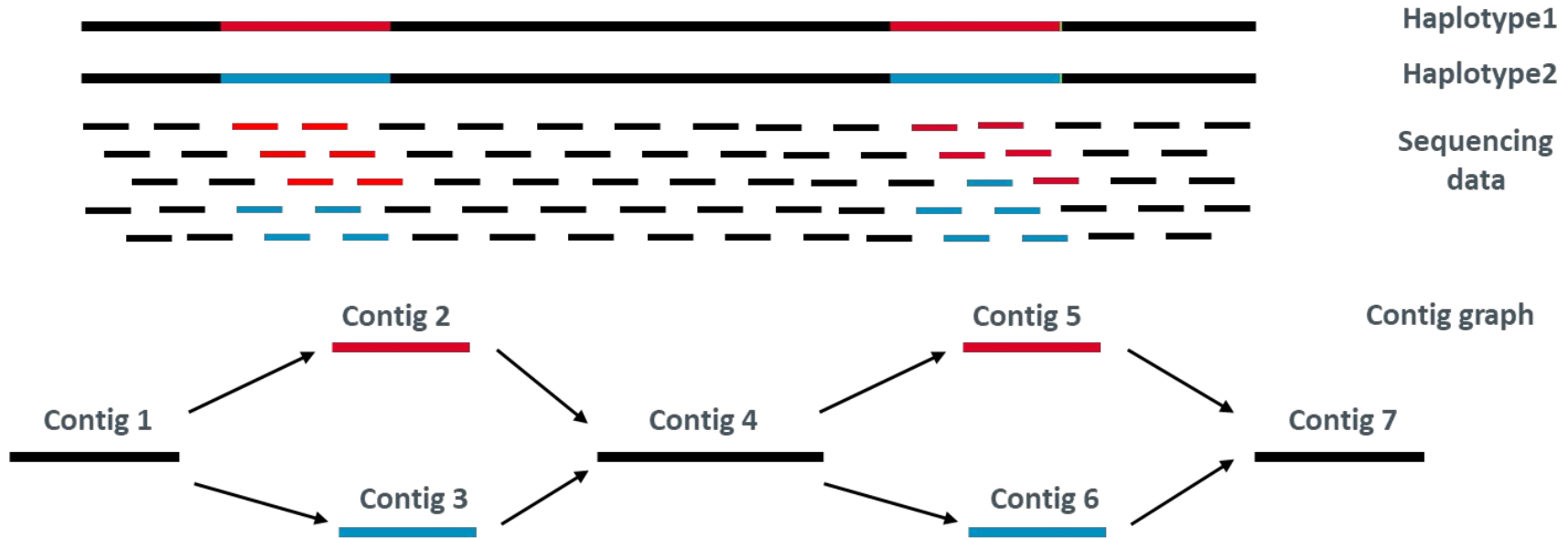
- Generally, they inspect the nucleotide one by one, and provide a correction for each nucleotide of the input assembly.
 - these algorithms are not able to properly process diploid genomes
=> switch from one haplotype to another



How polishing tools work.

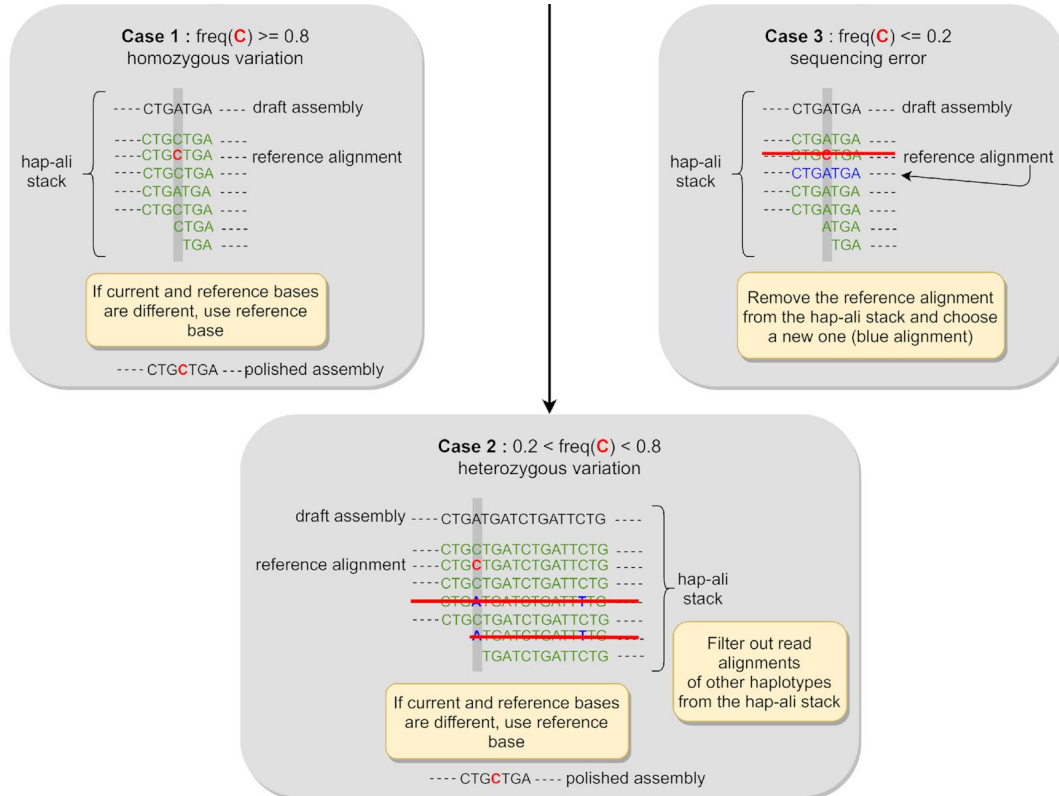


Genome assembly difficulties



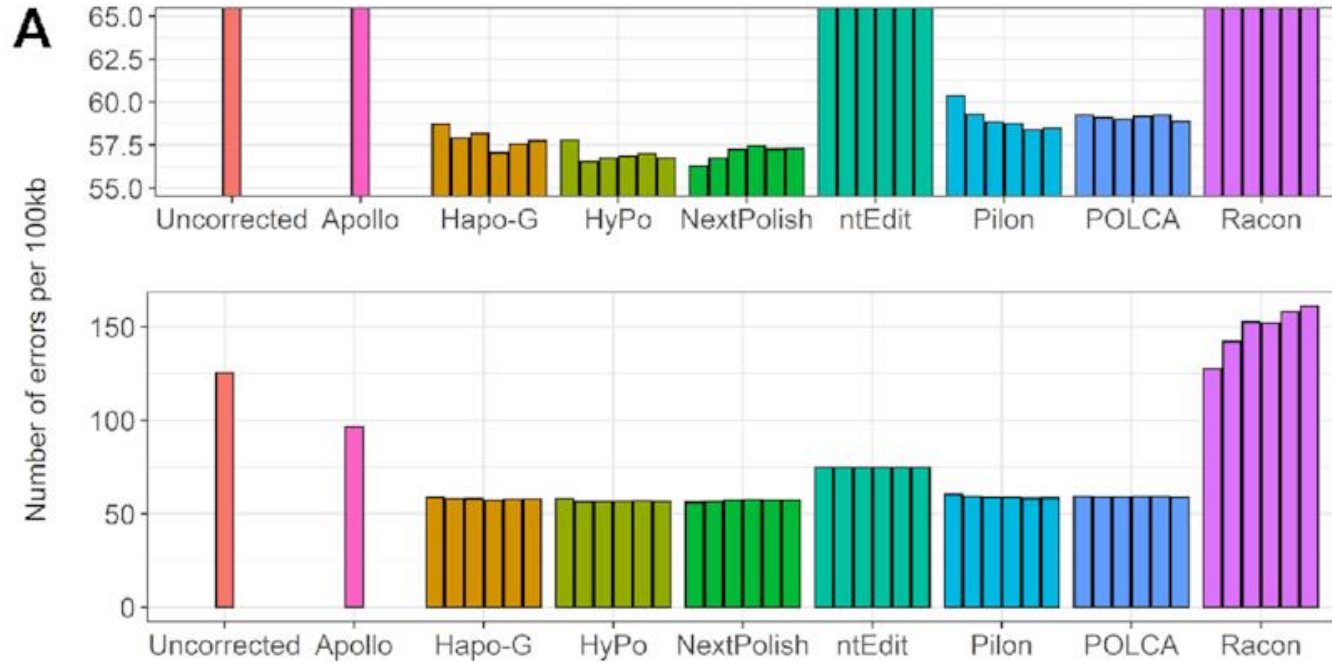
=> Heterozygous regions lead to fragmented assemblies and cause allelic duplication (over-estimate the size of the haploid genome)

How polishing tools work.



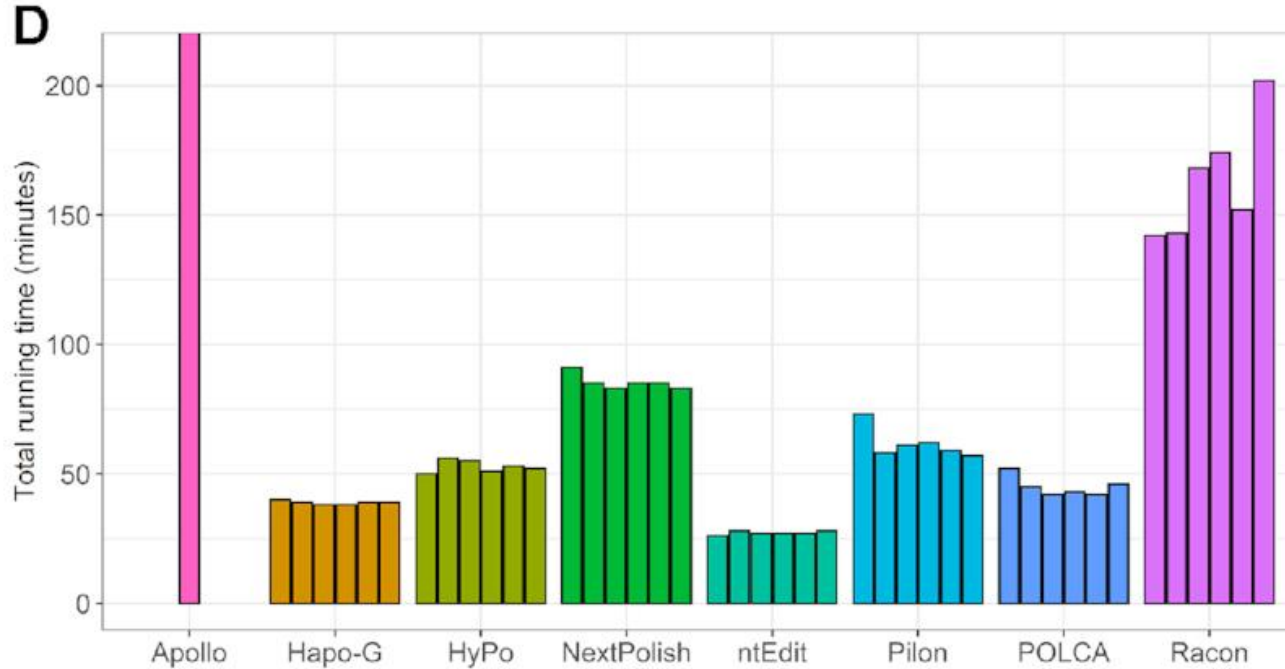
How polishing tools work.

Similar results on homozygous genome (*Arabidopsis thaliana*)



How polishing tools work.

Hapo-G is the faster among mapping-based methods



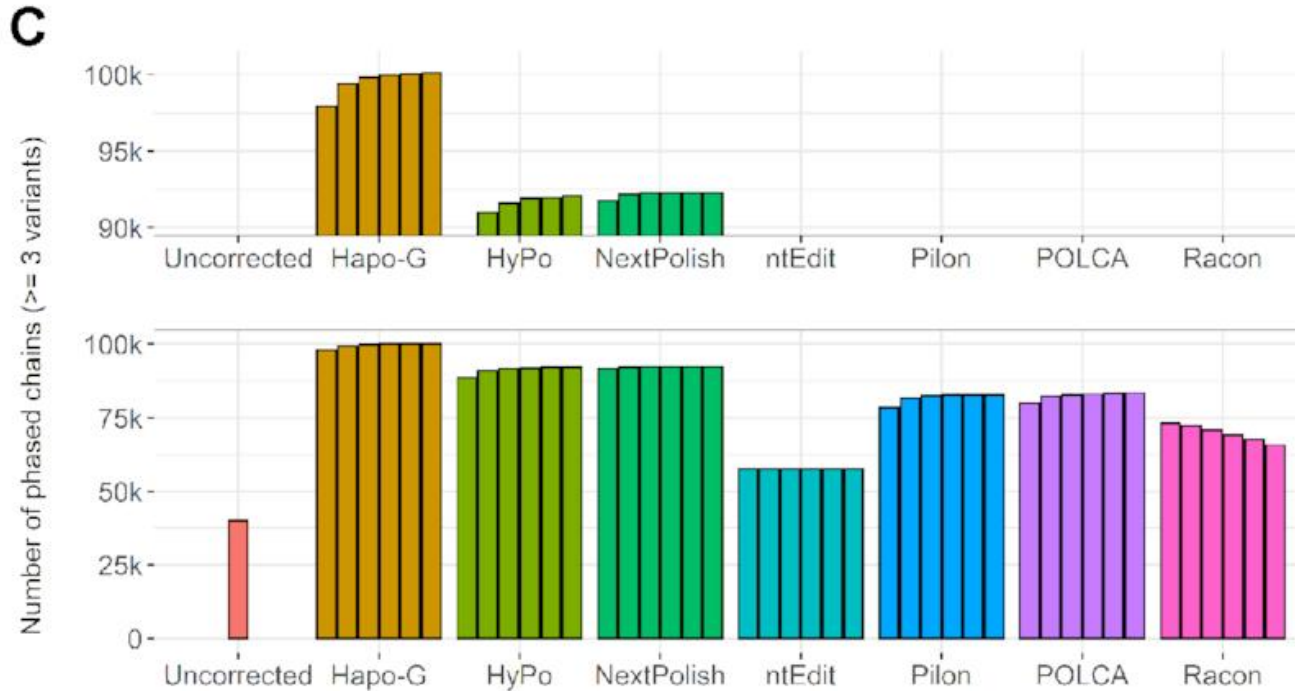
How polishing tools work.

Hapo-G generates less haplotype switches than other tools



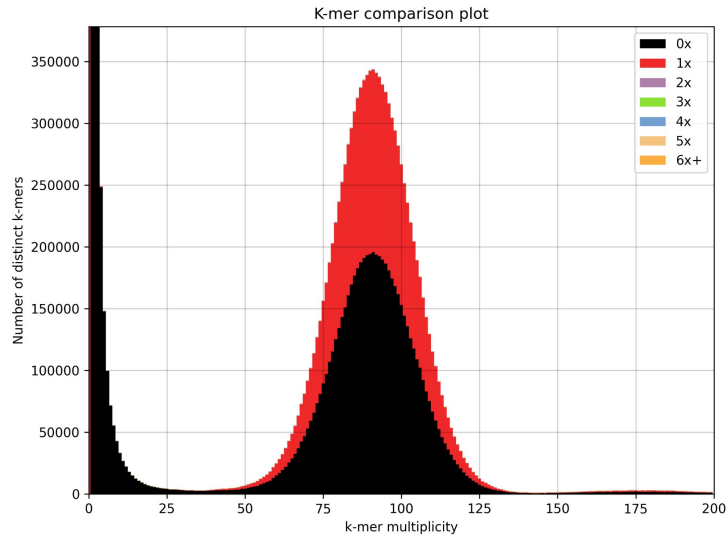
How polishing tools work.

Hapo-G generates less haplotype switches than other tools

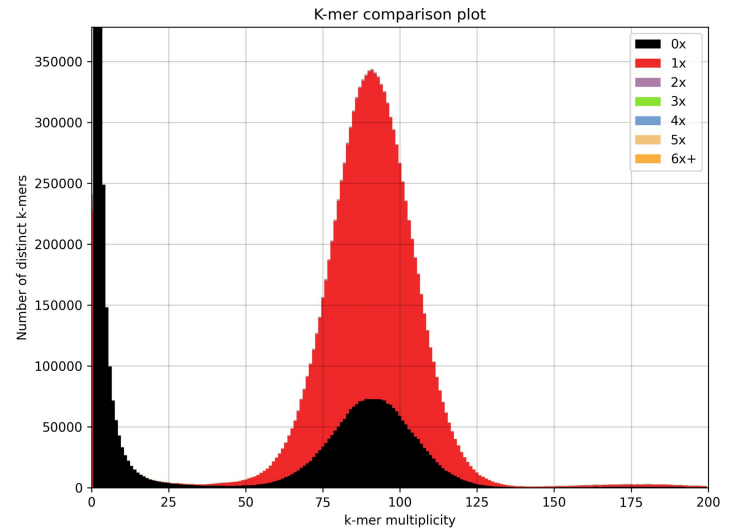


How to spot a potential problem with your assembly consensus.

- Each kmer of your readset should also be found in your genome assembly
=> Generate a KAT plot



=>



How to spot a potential problem with your assembly consensus.

- Errors in your consensus can affect gene prediction
=> Launch BUSCO and look at the “Complete”, “Fragmented” and “Missing” scores

***** Results: *****

```
C:57.3%[S:57.3%,D:0.0%],F:12.2%,M:30.5%,n:255
146      Complete BUSCOs (C)
146      Complete and single-copy BUSCOs (S)
0        Complete and duplicated BUSCOs (D)
31       Fragmented BUSCOs (F)
78       Missing BUSCOs (M)
255      Total BUSCO groups searched
```

***** Results: *****

```
C:75.3%[S:74.9%,D:0.4%],F:3.5%,M:21.2%,n:255
192      Complete BUSCOs (C)
191      Complete and single-copy BUSCOs (S)
1        Complete and duplicated BUSCOs (D)
9        Fragmented BUSCOs (F)
54       Missing BUSCOs (M)
255      Total BUSCO groups searched
```

Running a polishing in Galaxy

- Upload your genome assembly (fasta file) and data (usually two fastq files) or have access to it locally.
- Select the polishing tool (Hapo-G or Pilon) in the software package list (on the left).
- Select your dataset in the list
- Set parameters (usually first run with default)
- Hit the “execute” button

Running Hapo-G in usegalaxy.fr

Hapo-G genome polishing (Galaxy Version 1.3.3+galaxy0)

Genome assembly to polish

10: Polished assembly using Pilon

(--genome)

Type of data used for polishing

Short (paired) reads

Short (paired) reads collection

Long reads

Pre-aligned reads (BAM)

(--pe1)

Second set of short reads

3: SRR15597408_r2.fastq
2: SRR15597408_r1.fastq

(--pe2)

Include unpolished sequences in final output

No

(-u)

1

2

3

4

Conclusions

Polishing is needed, at least for genome assemblies based on error-prone reads

Check your assemblies (gene content, kat plot, merqury QV, ...)

Heterozygous regions are challenging, as most algorithms generate switches between haplotypes

Hands-on with *S. cerevisiae* (~12 Mb, 16 chromosomes)

Your mission is to perform, compare and give information about different assemblies :

- map your reads to the unpolished assembly using bwa-mem
 - use different polishing tools (Hapo-G, Pilon),
 - compare assemblies (Merqury QV, Busco)
-
- Processing will be performed using : <https://usegalaxy.fr/>
 - The data files are located at : Libraries / EBAll A&A 2022 / Polishing
 - You can access the flye assembly generated using ONT data (file flye_assembly.fasta)

Let's go!