From isolates to assemblies with PacBio

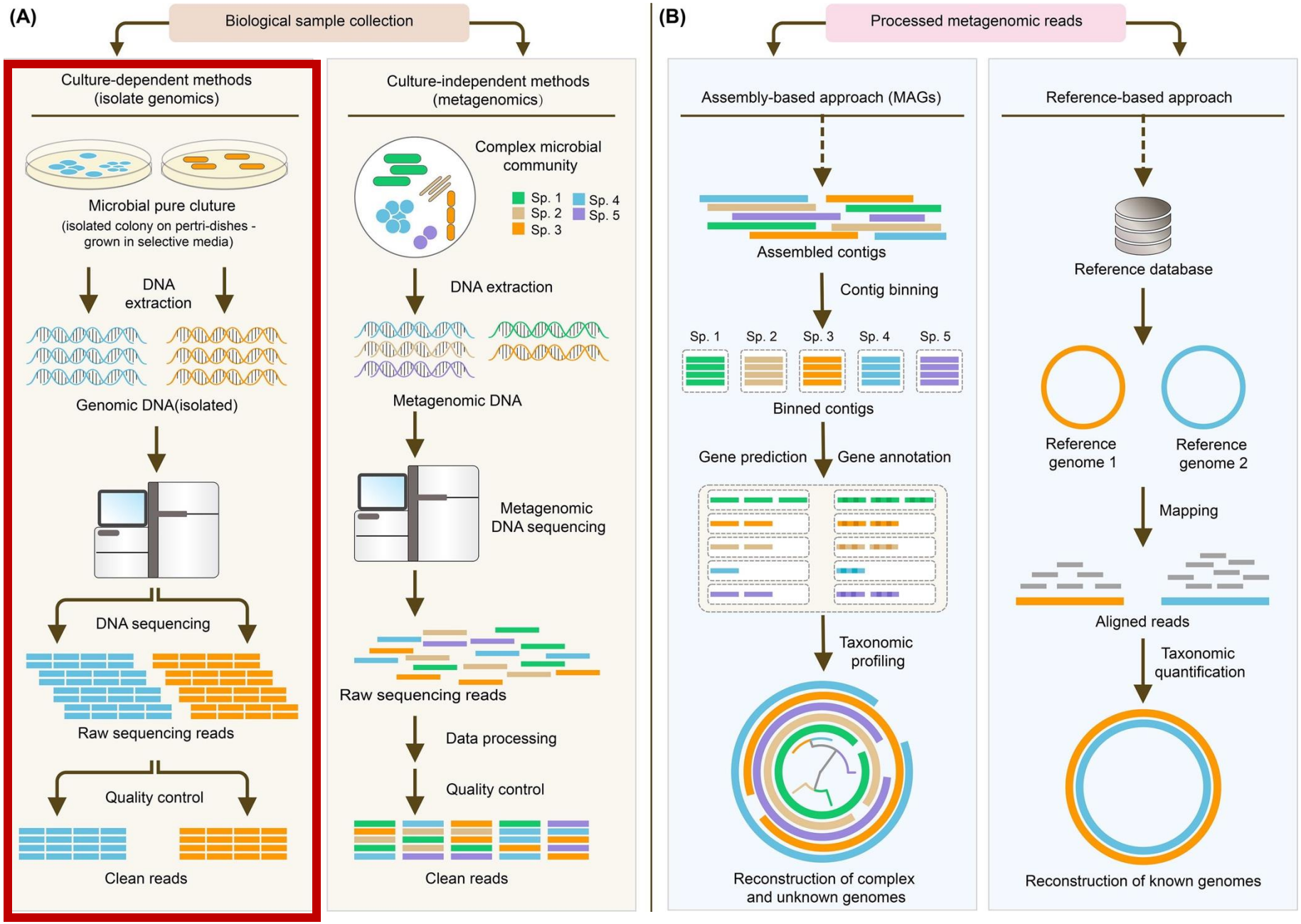
A pipeline of the Vonaesch Lab for the curnagl cluster

Simon Yersin

With the help of Julian Garneau, the SAGE team and Garance Sarton-Lohéac

Introduction

- This bioinformatic pipeline slides attempts to describe the steps to go from PacBio sequences of bacterial isolates to assembled genomes.



https://www.sciencedirect.com/science/article/pii/S2001037021004931

Isolations and DNA extraction

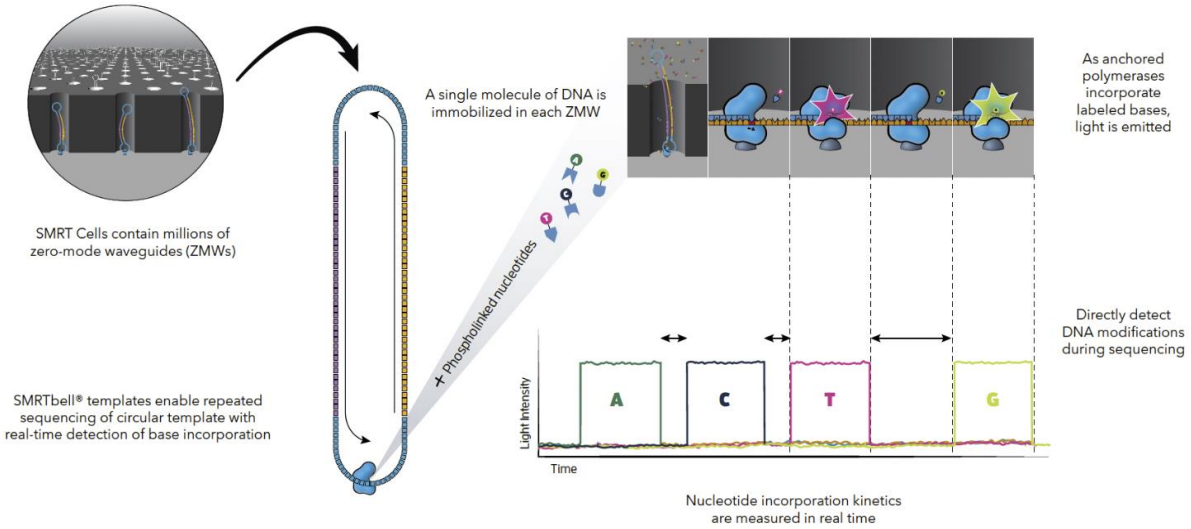
1. Isolation of bacteria
2. Grow in liquid broth overnight and perform DNA extraction using:
   1. Promega Wizard Genomic DNA Purification Kit
   2. Promega Maxwell RSC PureFood GMO and Authentication Kit
3. Prepare DNA samples for PacBio sequencing at the GTF facility
   1. Elute DNA in a TE buffer
   2. 500 ng of DNA in 25ul = 20ng/ul
   3. Prepare 4ul aliquot in PCR strip at 1ng/ul for DNA integrity analysis

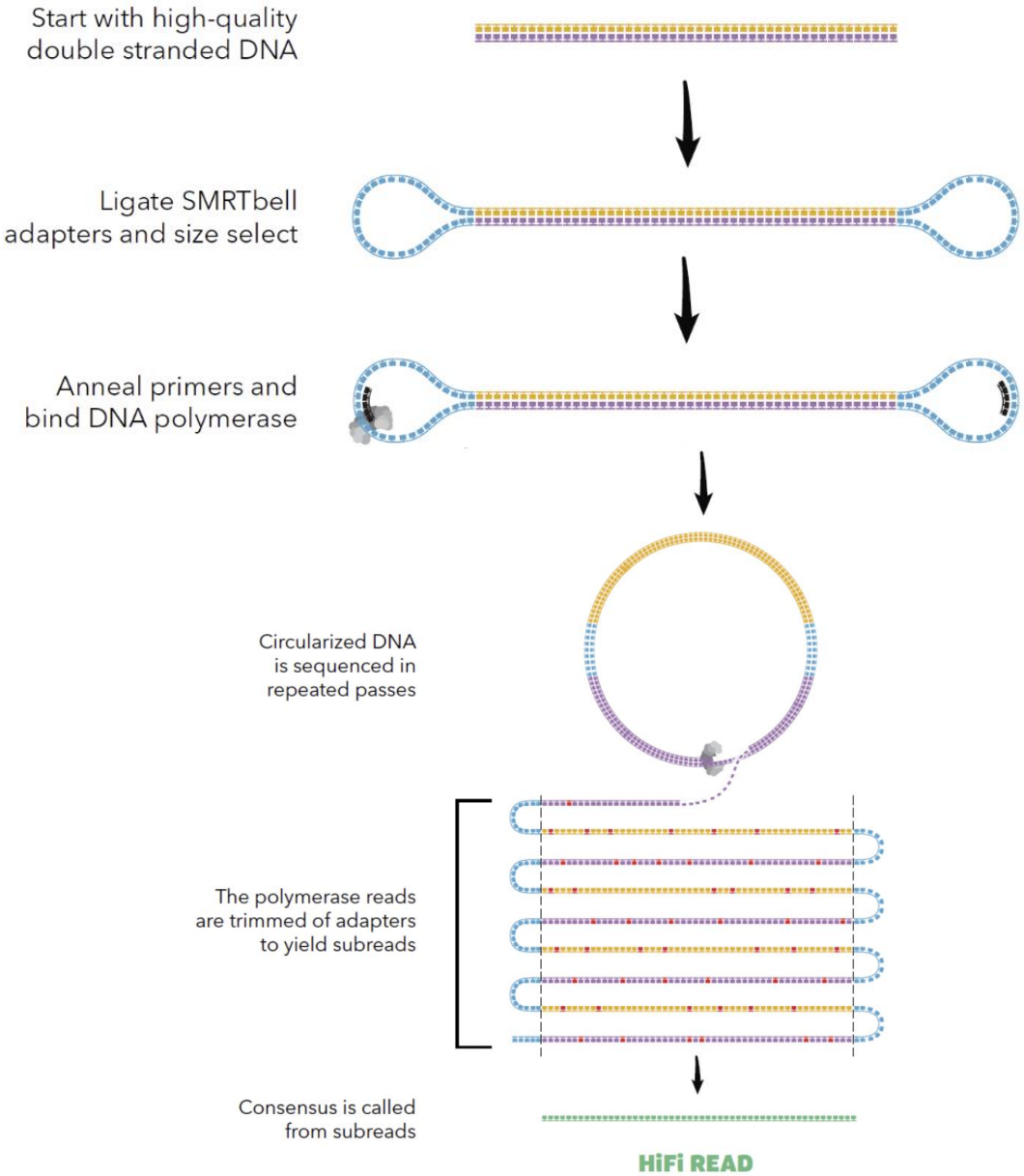
   https://wp.unil.ch/gtf/technology/

PacBio sequencing

- Main applications:
  - *De novo* genome sequencing
- HiFi long read sequencing:

They are loaded in a SMRT cell for Single Molecule Real Time sequencing. Sequencing movies can last for up to 30 hours.





PacBio sequencing

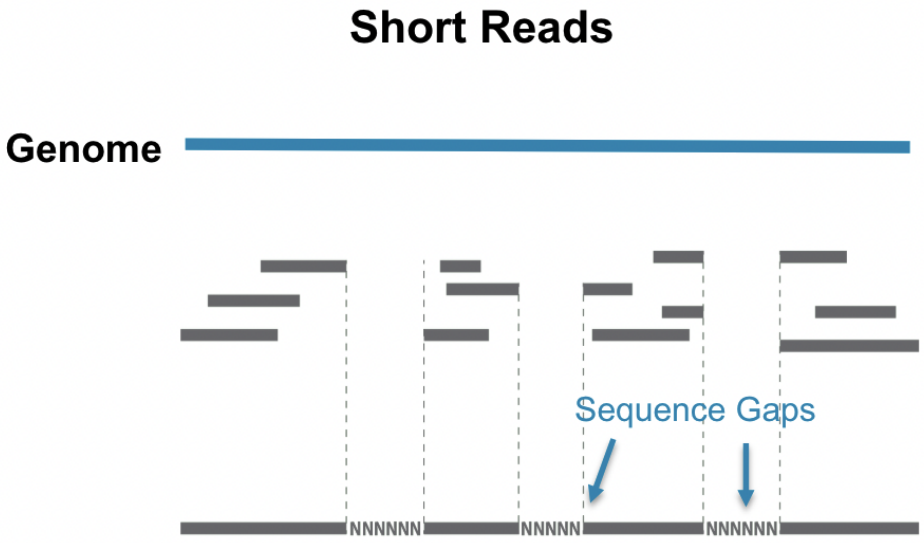
WHOLE GENOME SEQUENCING — HOW PACBIO COMPARES

	PacBio HiFi	Illumina	Oxford Nanopore
Average read length[1]	15–20 kb	2 x 150 bp	10–100 kb
Average read accuracy[1]	99.95% (Q33)	99.92% (Q31)	99.26% (Q21)
Coverage[2]	Unbiased	Reduced at low and high [GC]	Reduced in low-complexity runs
Variant calling: SNVs	✓	✓	✓
Variant calling: indels	✓	✓	✗
Variant calling: SVs	✓	✗	✓
Genome assembly: contiguity	✓	✗	✓
Genome assembly: accuracy	✓	✓	✗
Epigenetics: 5mC	✓	✗	✓

1. PacBio HiFi: HG003 18 kb library, Sequel II system chemistry 2.0, precisionFDA *Truth Challenge* V2 (https://doi.org/10.1101/2020.11.13.380741), Illumina: HG002 2×150 bp NovaSeq library, precisionFDA *Truth Challenge* V2 (https://doi.org/10.1101/2020.11.13.380741), ONT: Q20+ chemistry (R10.4, Kit 12), Oct 2021 GM24385 Q20+ Simplex Dataset Release (https://labs.epi2me.io/gm24385_q20_2021.10/)

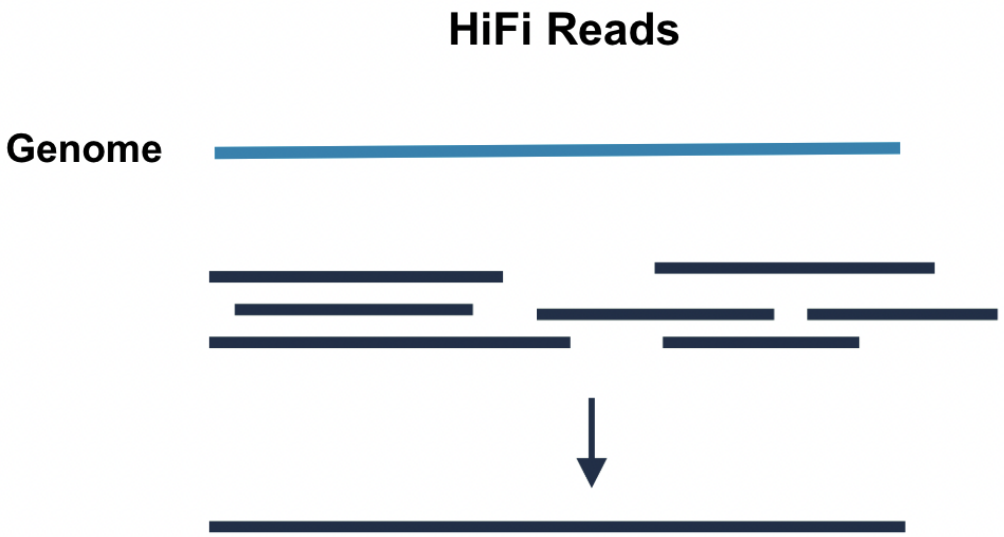
2. HiFi+ONT: Nurk 2021 https://doi.org/10.1101/2021.05.26.445798, HiFi+Illumina: Logsdon 2020 https://doi.org/10.1038/s41576-020-0236-x, ONT: Tan 2022 https://doi.org/10.1101/2022.01.11.475254

DRAFT VS COMPLETE GENOME ASSEMBLY



Draft Genome

Missing sequencing leads to missed genes and limits biological interpretation



Complete Genome

A comprehensive structural, functional and organizational picture of the genome

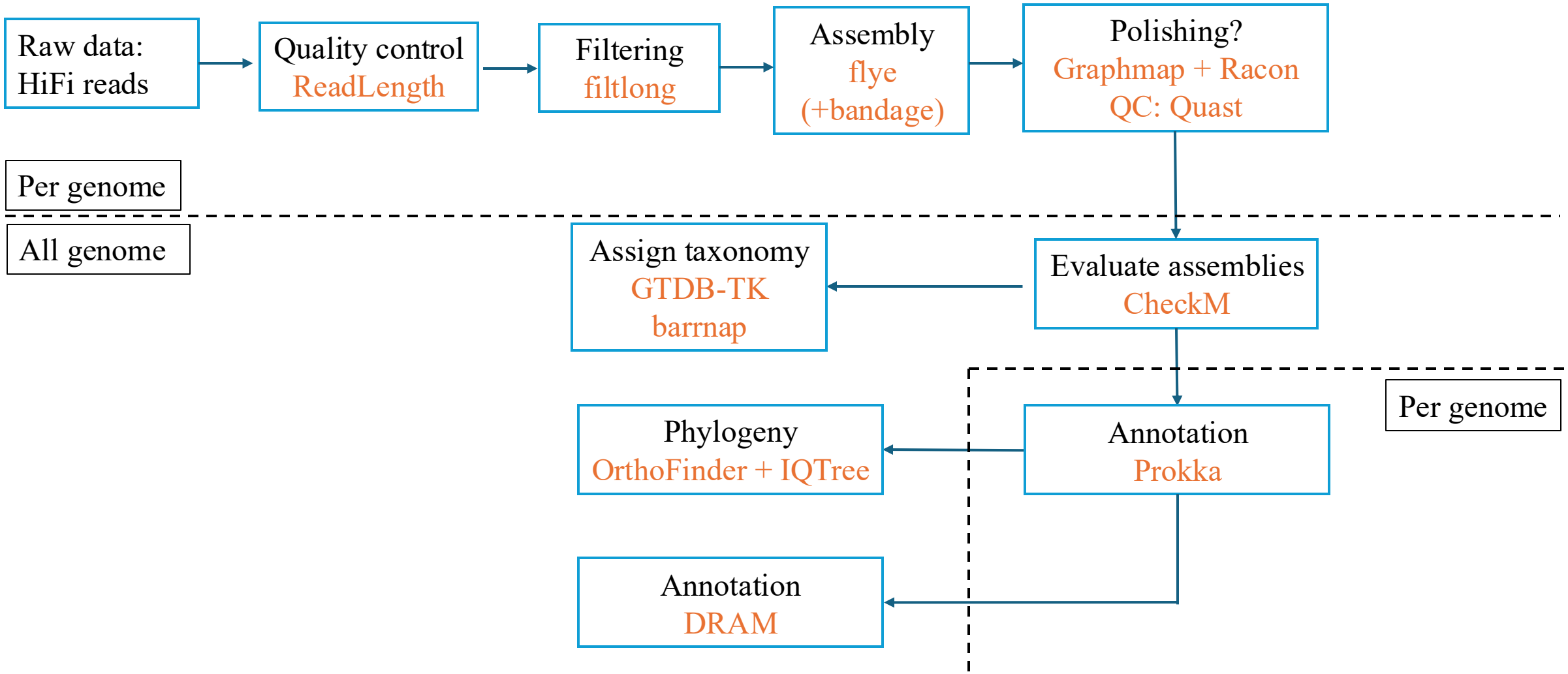
Metrics for high quality genome

- Compute: HiFi assemblies are faster than traditional long read assemblies
- Contiguity: we want fewer contigs and long contigs for a continuous assembly
- Completeness: total size
- Correctness: Proportion of the assembly that is free of mistakes

N50: size of the shortest contig covering 50% of the total assembly → we want higher

L50: number of contig making up 50% of the cumulated contig length → we want small

General pipeline



STEP 0: PREPARATION

- Prepare your working directory with sub-directories:
  - S_salivarius_pacbio
    - Raw_data
    - Scripts
    - Outputs
- Then copy your raw data from the nas to Raw_data folder:

```
cp /nas/FAC/FBM/DMF/pvonaesc/default/D2c/Simon\ Yersin/<raw data directory> ~/S_salivarius_pacbio/Raw_data
```

- Wait for the copy to finish and verify with ls that all your files are copied
- In the terminal, run:
  - module purge
  - dcsrsoft use arolle

STEP 1: STATISTICS AND QC

- Extract read length of your fastq files using:
  - 01_qc.sh
- Download the readLength.txt files on your local computer
- Switch to RStudio and follow the script:
  - ReadLength_PacBio.Rmd
    - Save the histogram, the cumulative summary and the calculations
- Adapt and use read_count.sh to save a file with the number of reads before filtering
  - read_count.sh

STEP 2: FILTERING

- Filter your reads using *Filtlong*:
  - 02_filtlong.sh
  - Adapt the variables:
    - MINIMUM_read_LENGTH
    - Min_mean_q
    - Length_weight
    - Target_bases
    - Variable with directory + array
- Adapt and use read_count.sh to save a file with the number of reads after filtering
  - read_count.sh

STEP 3: ASSEMBLY

- Assemble your genome using *flye*:
  - 03_flye_assembler.sh
  - Adapt variables and array
  - Follow the progress of the job with Squeue, sacct or by opening the log file of the assembly
- Download assembly graph (assembly_graph.gfa) and open it on Bandage
  - File > load graph
  - Draw graph
  - Save assembly graph

STEP 4-5: POLISHING & QC

Polishing allows to remove errors such as SNVs and INDELS. Usually, a first round of polishing is done with the long-reads assemblies then a second round with the short-reads (we do not have short reads in our case)

- Decompress filtered read files

  gzip –d *.fatsq.gz

- Polish the assembled genome using *Graphmap* and *Racon*:

  04_polishing.sh

- Inspect the quality of your polishing using *Quast*:

  05_qc_polishing.sh

  Download report.html to visualize the QC

STEP 6: Evaluate the assemblies

- Copy final assemblies into new folder: comp_genomics/genomes
- This script uses the conda environment checkm. It is activated at the start of the script
- Evaluate the quality of your assemblies using *CheckM* :
  06_checkM.sh
- Download report: CheckM_QC_stats.tsv

STEP 7: GTDB-TK

- GTDB-TK is a software toolkit for assigning objective taxonomic classifications to bacterial and archaeal genomes based on the Genome Database Taxonomy.
- This script uses the conda environment gtdbtk. It is activated at the start of the script.
- Run GTDB-TK to obtain the taxonomic classification of your genomes, using:

  07_gtdb_tk.sh
- Download the output file: gtdbtk.bac120.summary.tsv

Downstream steps and analysis

- From here we have assembled, filtered, QC controlled, and taxonomy assigned genomes from Illumina short reads
- The next steps depend on the research question, here are some options:
  - Genes annotation and comparative genomic (Anv'io, discuss with Julian)
  - Strain diversity analysis using fastANI or inStrain
  - Phylogeny

FINAL NOTES

- The standard output files .err and .out can be erased once runs are completed without failures or errors
- Download all scripts files or save them on the NAS for reproducibility
- Download output files needed for downstream analysis
- Clean conda with: conda clean –tarballs –packages

Good job!