Preprocessing of metagenomic sequencing reads

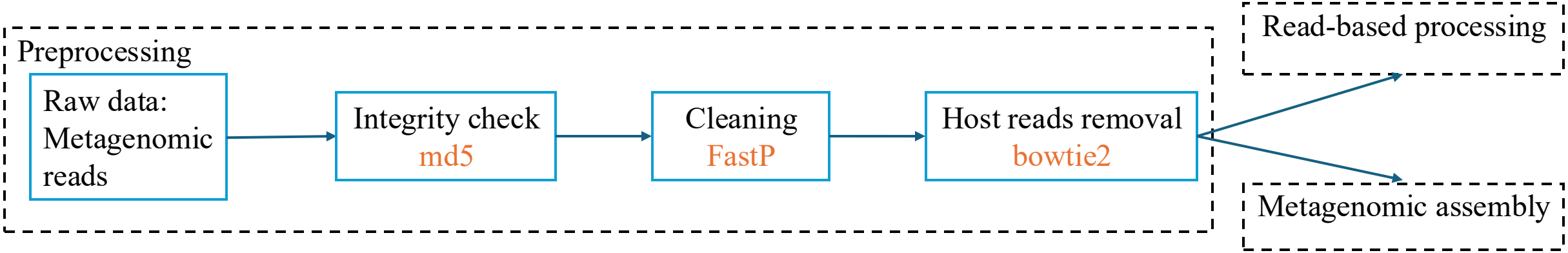
A pipeline of the Vonaesch Lab for the curnagl cluster

Simon Yersin, Julian Garneau, Margaux Crézé

Introduction

- This bioinformatics scripts are meant to be used on the HOC cluster, either Urblauna (for sensitive data typical containing human reads) or Curnagl.
- The pipeline include the following steps:
  - Integrity check
  - Merging duplicated reads
  - Cleaning
  - Host reads removal

Pre-processing steps



STEP 0: Preprocessing

- We are assuming that the following steps have already been performed upon reception of the data:
  - Storage on the NAS and a hard drive
- Check the integrity of the data and merge/rename the reads files using:
  - 0_preprocessing.sh

STEP 1: Cleaning

- Clean adapters and low-quality reads using FastP:
  - 01_fastp.sh

STEP 3: Host reads removal

- Remove host reads (human reads) using bowtie2 aligner and a reference genome. For human, see the bowtie2 index.
  - 📜 01_2_host_reads.sh (adapt index!)

- Clean reads can be then used for either metagenomic assembly or read based analysis (taxonomic assignation, functional potential)