Introduction to Metagenomics for Clinical Virology

Sarah Buddle

UCL Great Ormond Street Institute of Child Health

Session developed by Dr Cristina Venturini

Session structure

15:30-16:30: Introduction to metagenomics

16:30-18:00: Metagenomics bioinformatics practical

1. What is metagenomics?

2. What clinical questions can we answer with metagenomics?

3. What are the advantages and disadvantages of metagenomics over other techniques you might use to answer those questions?

4. (Optional) What might you need to consider before implementing metagenomics in a clinical or public health setting? If you have used metagenomics before, what difficulties did you encounter?

1. What is metagenomics?

- Sequencing all the genetic material in a sample
- Not targeting to one or a small number of organisms
- In context of viruses, sequencing DNA and RNA

2. What clinical questions can we answer with metagenomics?

- What pathogens are there?
  - What is causing the disease?
  - What is the composition of the microbial community?
  - Surveillance: Are there any novel strains or species?

- What are the genome sequences of the viruses?
  - Antiviral resistance
  - Tracking of outbreaks

3. What are the advantages and disadvantages of metagenomics over other techniques you might use to answer those questions?

- Advantages
  - No prior assumptions – good for new or unusual organisms
  - Sequence information

- Disadvantages
  - Contamination
  - Expensive and time consuming
  - Lots of infrastructure and trained staff required
  - Can be less sensitive than PCR/large inputs required
  - Regulatory and accreditation challenges

4. (Optional) What might you need to consider before implementing metagenomics in a clinical or public health setting? If you have used metagenomics before, what difficulties did you encounter?

Protocol

What are the key steps in a metagenomics protocol?

What is the purpose of each step?

What methods might you use?

Sequencing

QC

Sample collection

Host removal

Assembly

DNA/RNA extraction

Library prep

Classification

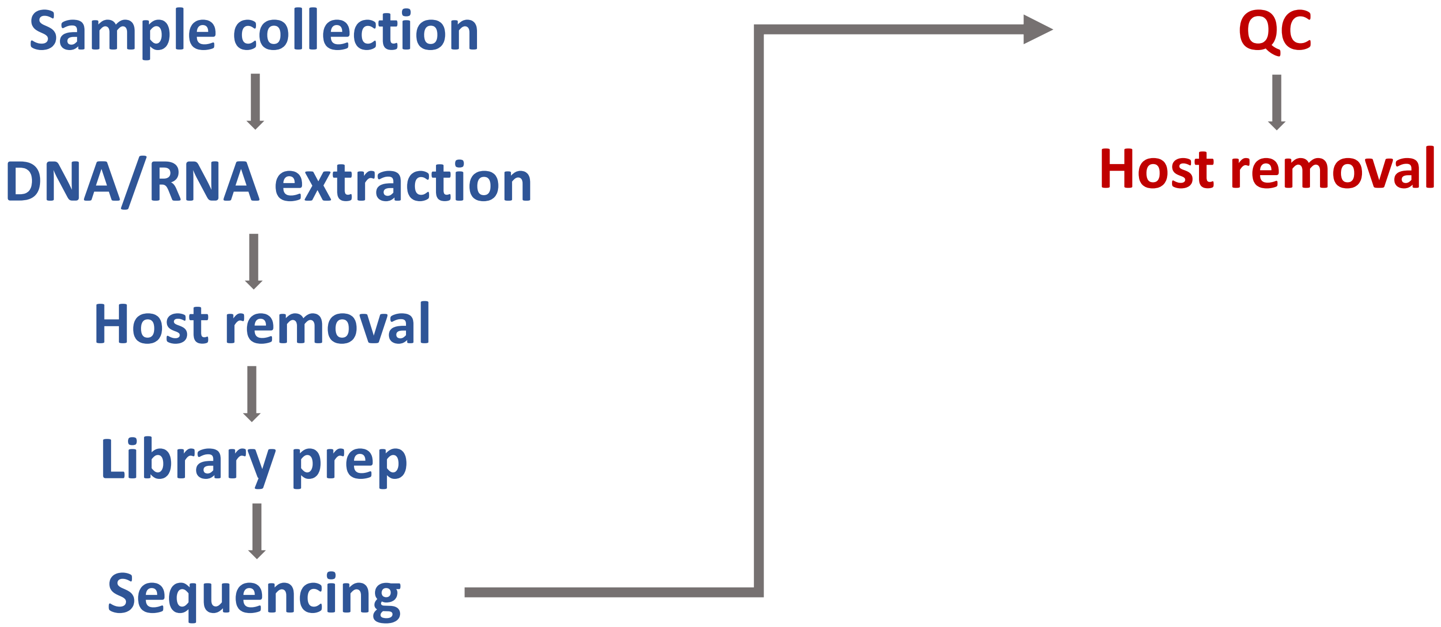
Host removal

Viral sequence analysis

Reporting

Optional: What sequencing platforms could you use for metagenomics and what are the advantages/disadvantages of each?

Protocol



Sample collection

DNA/RNA extraction

Host removal

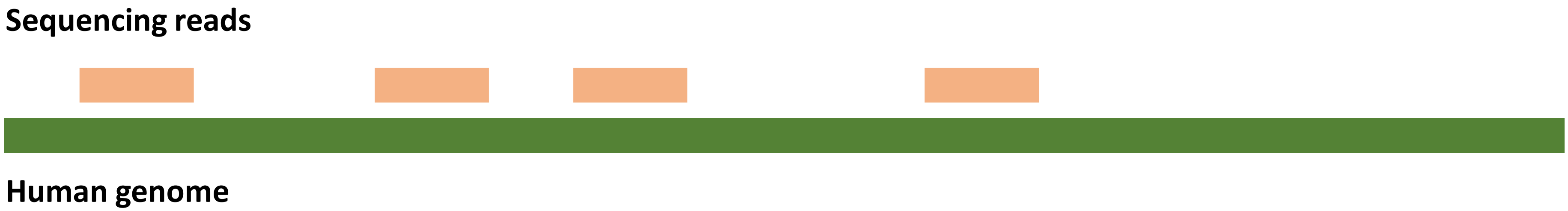
Library prep

Sequencing

QC

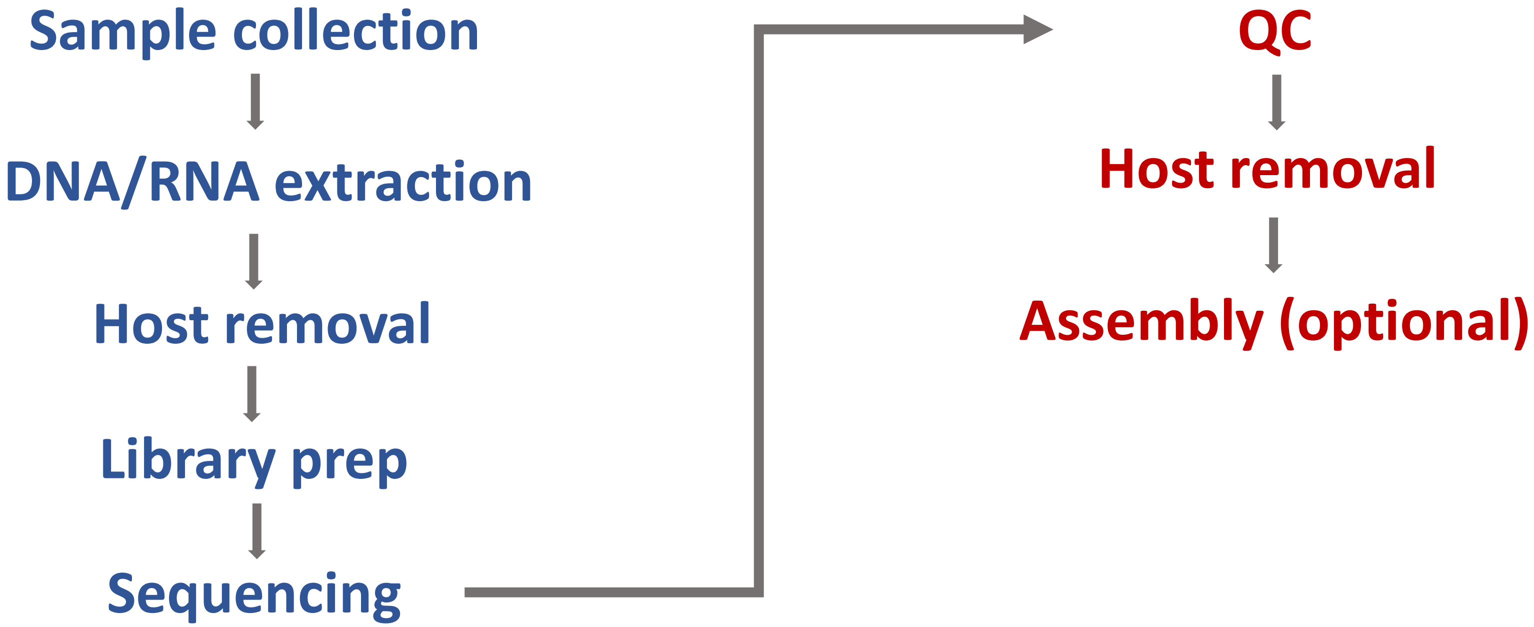
Host removal

Host removal: alignment



Preliminary round with a quick classifier also an option

Protocol



Sample collection

DNA/RNA extraction

Host removal

Library prep

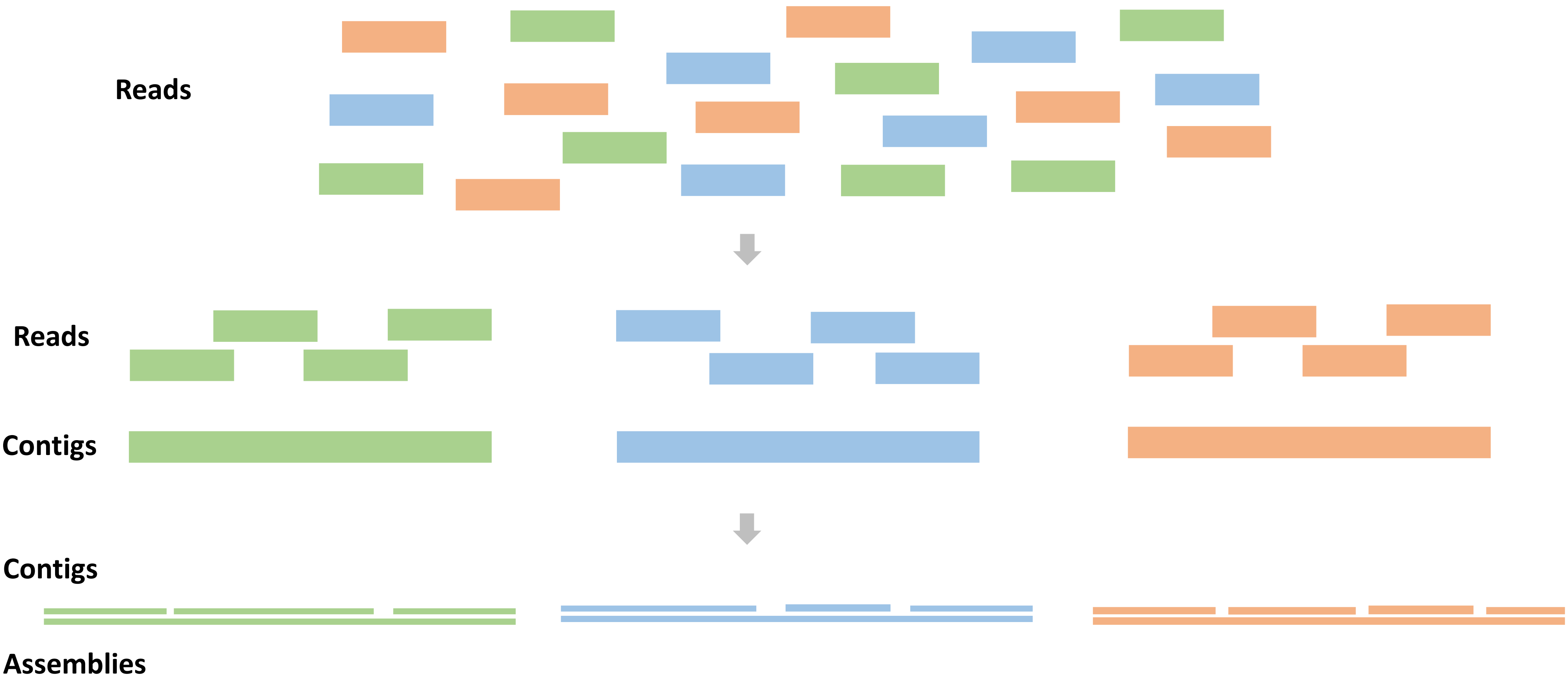
Sequencing

QC

Host removal

Assembly (optional)

Assembly



Reads

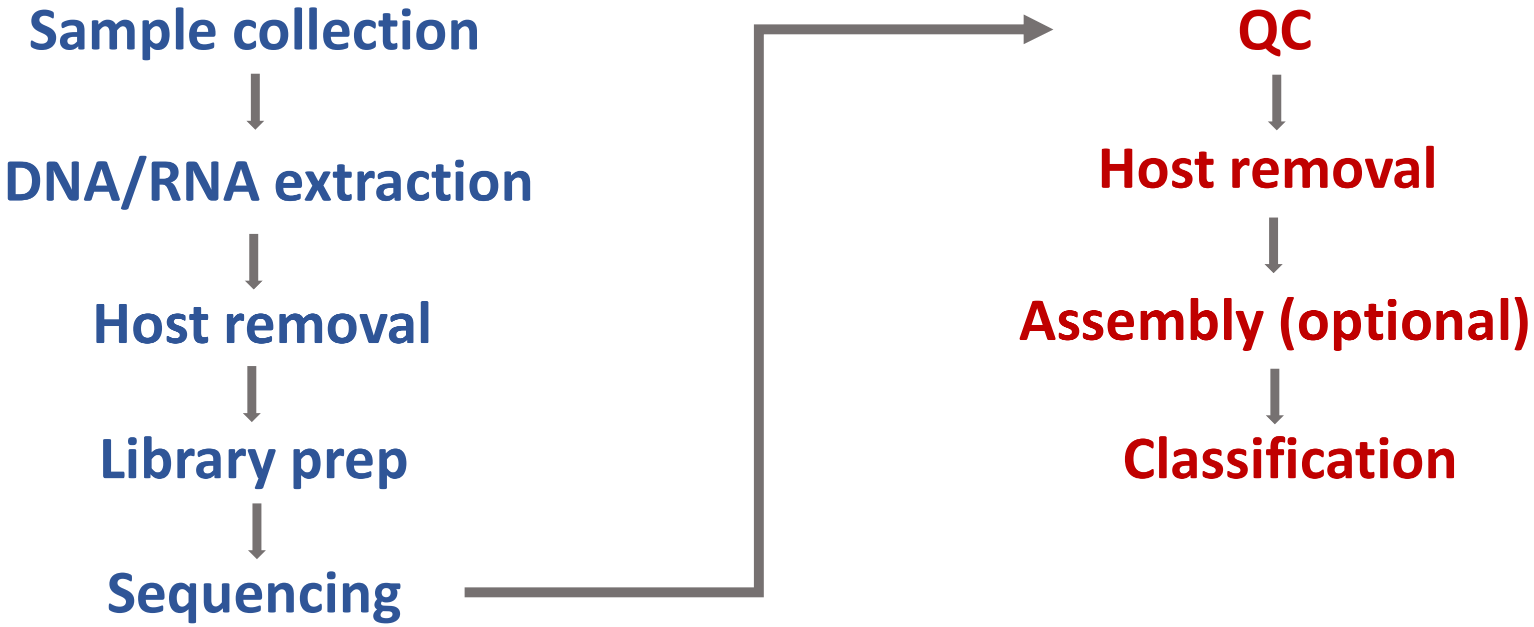
Reads

Contigs

Contigs

Assemblies

Protocol



Sample collection

DNA/RNA extraction

Host removal

Library prep

Sequencing

QC

Host removal

Assembly (optional)

Classification

Classification

Classification is deciding which species (or other taxonomic group) a read corresponds to

Reads are classified by comparison to a reference database containing known genome sequences

Challenge: some parts of DNA are similar in different organisms

Classification tools

Alignment-based
E.g. BLAST, DIAMOND

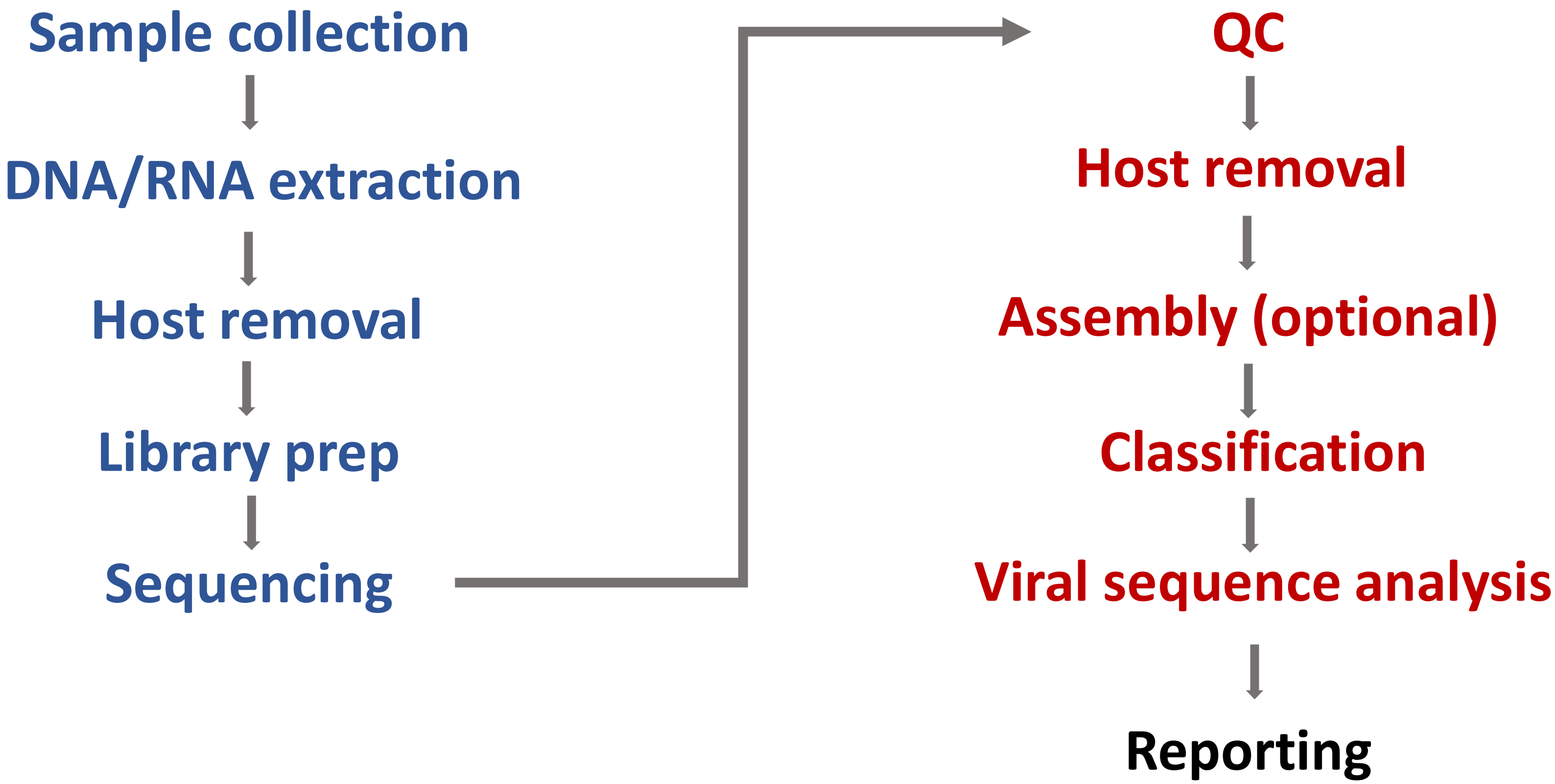
K-mer-based
E.g. Kraken2, Centrifuge

Marker gene-based
E.g. mOTU, MetaPhlAn

Nucleotide-based
E.g. BLASTN

Protein-based
E.g. DIAMOND, Kaiju

Protocol



Sample collection

DNA/RNA extraction

Host removal

Library prep

Sequencing

QC

Host removal

Assembly (optional)

Classification

Viral sequence analysis

Reporting

Optional: What sequencing platforms could you use for metagenomics and what are the advantages/disadvantages of each?

Classification

What factors should we consider when choosing:

1: a classifier

2: sequences to include in your database

Classification

How should we choose a classifier?

- Suitability for type of sequencing and microbe
- Sensitivity and specificity
- Time and computational resource requirements
- Ease of use

Classification

How should we choose a database?

- What organisms to include
- Nucleotide vs protein (protein good for more divergent viruses but can give more false positives)
- Prebuilt vs custom

Contamination

1. Where might contamination come from?

2. How can we reduce/deal with contamination?

Contamination

Where might contamination come from?

- From the patient (e.g. skin flora)
- Lab contaminants
- Kitome (microbes present in reagents etc.)
- Index hopping
- Bioinformatic contaminants – misclassification

Contamination

How can we reduce/deal with contamination?

- Sterile environment in lab
- Negative controls
- Database choice
- Quality control and thresholds

Practical

Part 1: Metagenomics analysis with Kraken2/Bracken (command line)

Try to work out the commands yourself rather than looking at the answers!

Part 2: Metagenomics analysis with CZID (online)

Use the login details on the board.

Commands to recap

Less (view file)

>   (Redirect to file)

Command --help /   man *command*   (view manual)

Choosing bioinformatics protocols for metagenomics

The protocol shown in the practical may not the best one for your research or clinical question!

Some other tools: a non-exhaustive list

nf-core/taxprofiler

nf-core is a set of community-curated best practice bioinformatics pipelines built in Nextflow.
Taxprofiler Includes Kraken2/Bracken, DIAMOND, Centrifuge etc

CZ ID

Online, cloud-based, user-friendly tool

Illumina Dragen Metagenomics / Nanopore EPI2ME labs wf-metagenomics
Illumina and Nanopore's tools. Simple to run and can be automated.

EPI2ME Labs

Check benchmarking papers for lots of other options!