Introduction to Metagenomics for Clinical Virology

Sarah Buddle
UCL Great Ormond Street Institute of Child Health

Session developed by Dr Cristina Venturini

Session structure

11:00-12:00: Introduction to metagenomics

12:00-13:00: 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?

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

Protocol

What are the key steps in a metagenomics protocol?

What is the purpose of each step?

What methods might you use?

Sequencing

Sample collection

QC

Host removal

Assembly

Library prep

DNA/RNA extraction

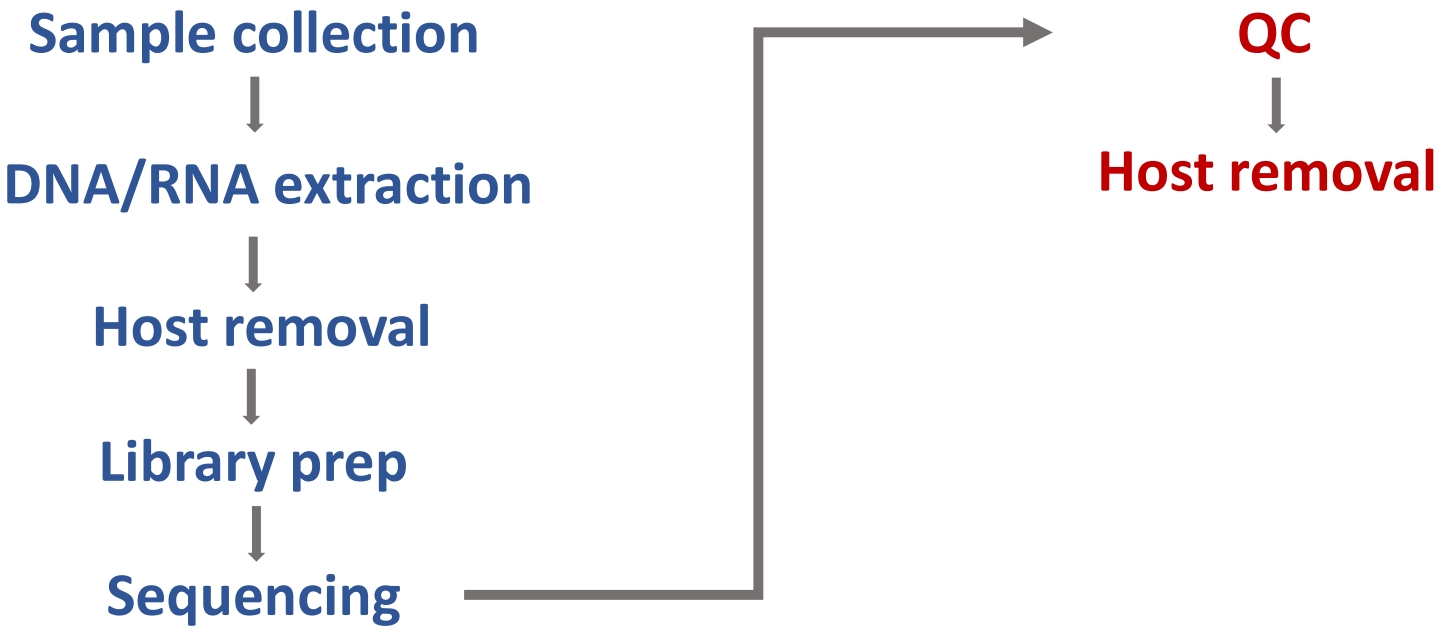
Classification

Host removal

Viral sequence analysis

Reporting

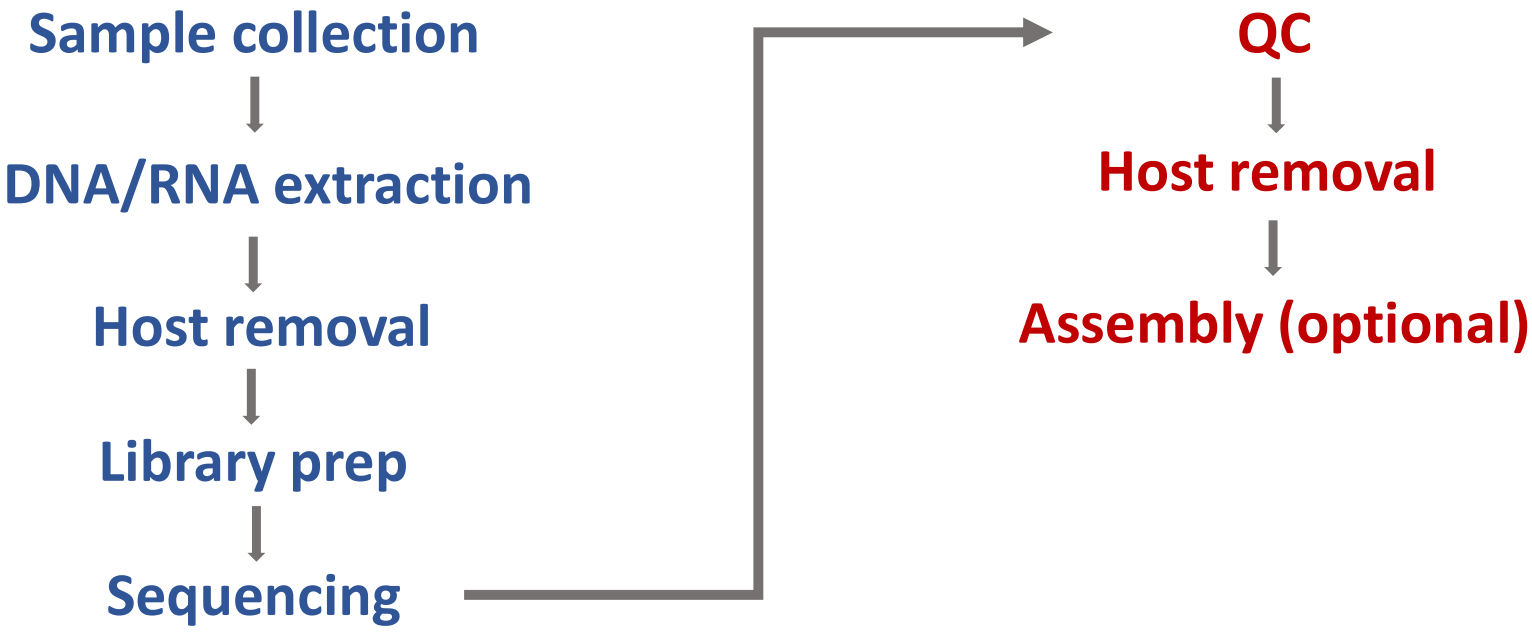
Protocol


Sample collection
DNA/RNA extraction
Host removal
Library prep
Sequencing
QC
Host removal

Host removal: alignment

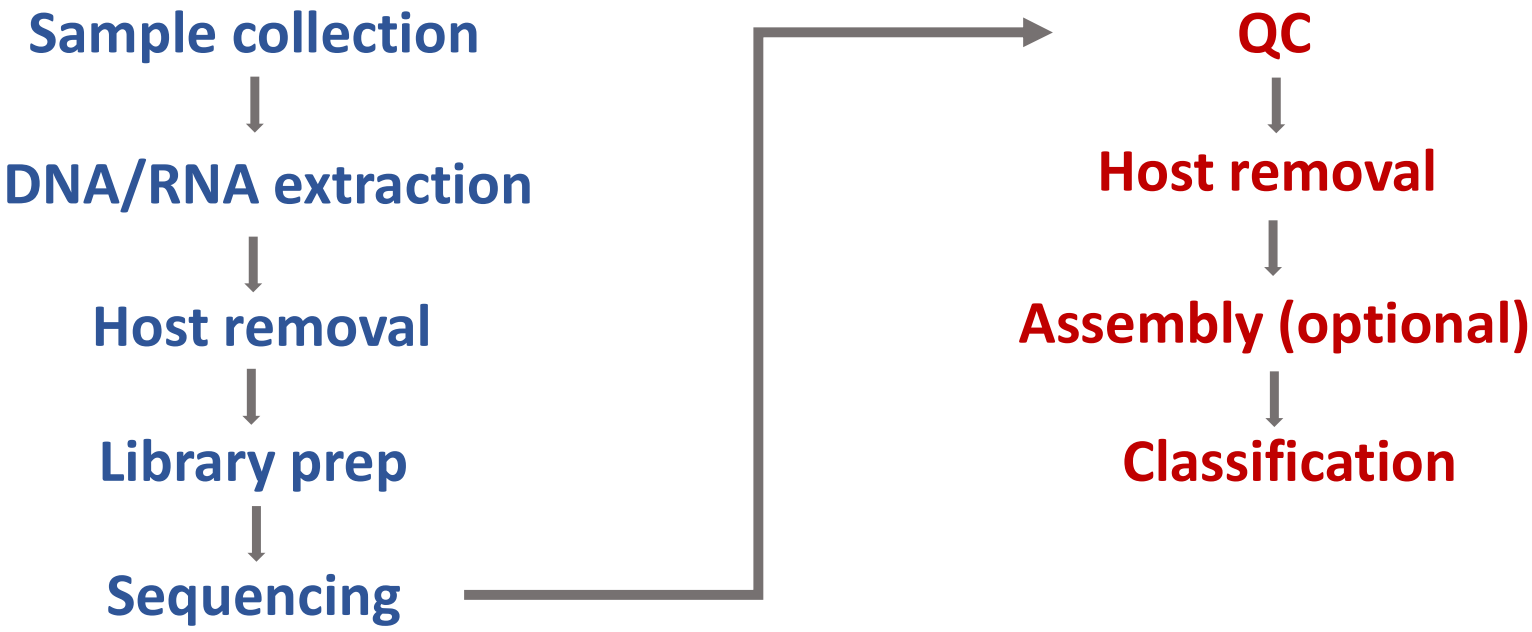
Preliminary round with a quick classifier also an option

Protocol


Sample collection
DNA/RNA extraction
Host removal
Library prep
Sequencing
QC
Host removal
Assembly (optional)

Assembly

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 databases containing known genome sequences

Challenge: some parts of DNA are similar in different organisms

Classification tools

Alignment-based
E.g. BLAST, DIAMOND

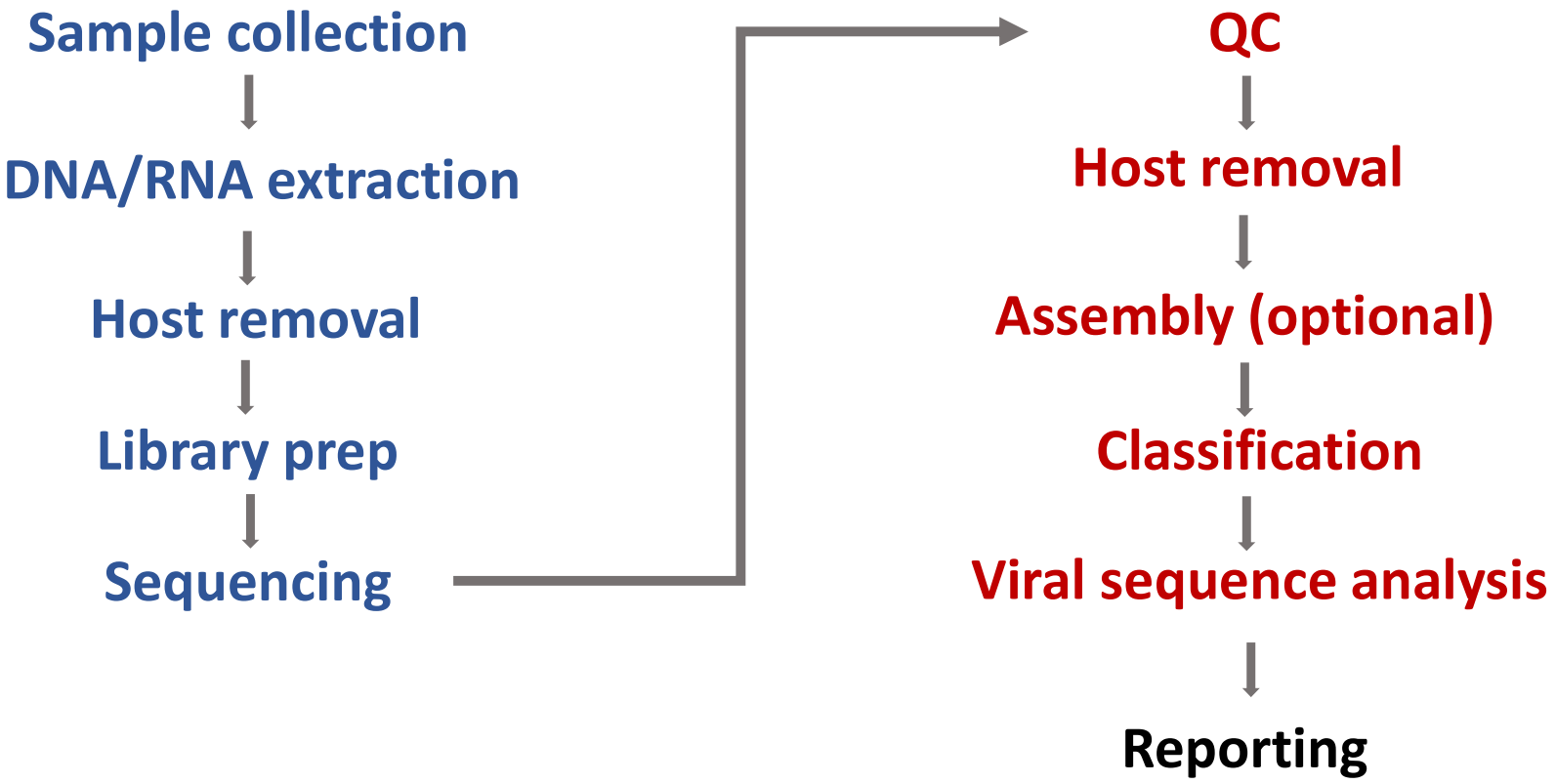
K-mer-based
E.g. Kraken2, Centrifuge

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

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
- 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: Metagenomics for diagnosis

Background:

We'll be analysing a clinical specimen from a 41-year-old patient who reported fever, chest tightness, cough, pain and weakness, and was admitted to hospital 6 days after the onset of the disease. The patient with no history of hepatitis, tuberculosis or diabetes. Preliminary investigations excluded the presence of influenza virus, *Chlamydia pneumoniae* and *Mycoplasma pneumoniae* and this was confirmed by PCR. Other common respiratory pathogens, including human adenoviruses, also tested negative. To investigate other possible infectious causes of disease, bronchoalveolar lavage fluid (BALF) was collected and deep metatranscriptomic sequencing was performed.

Task:

You need to analyse the data to determine the cause of the disease and determine what treatment options may be available.

Practical: Metagenomics for diagnosis

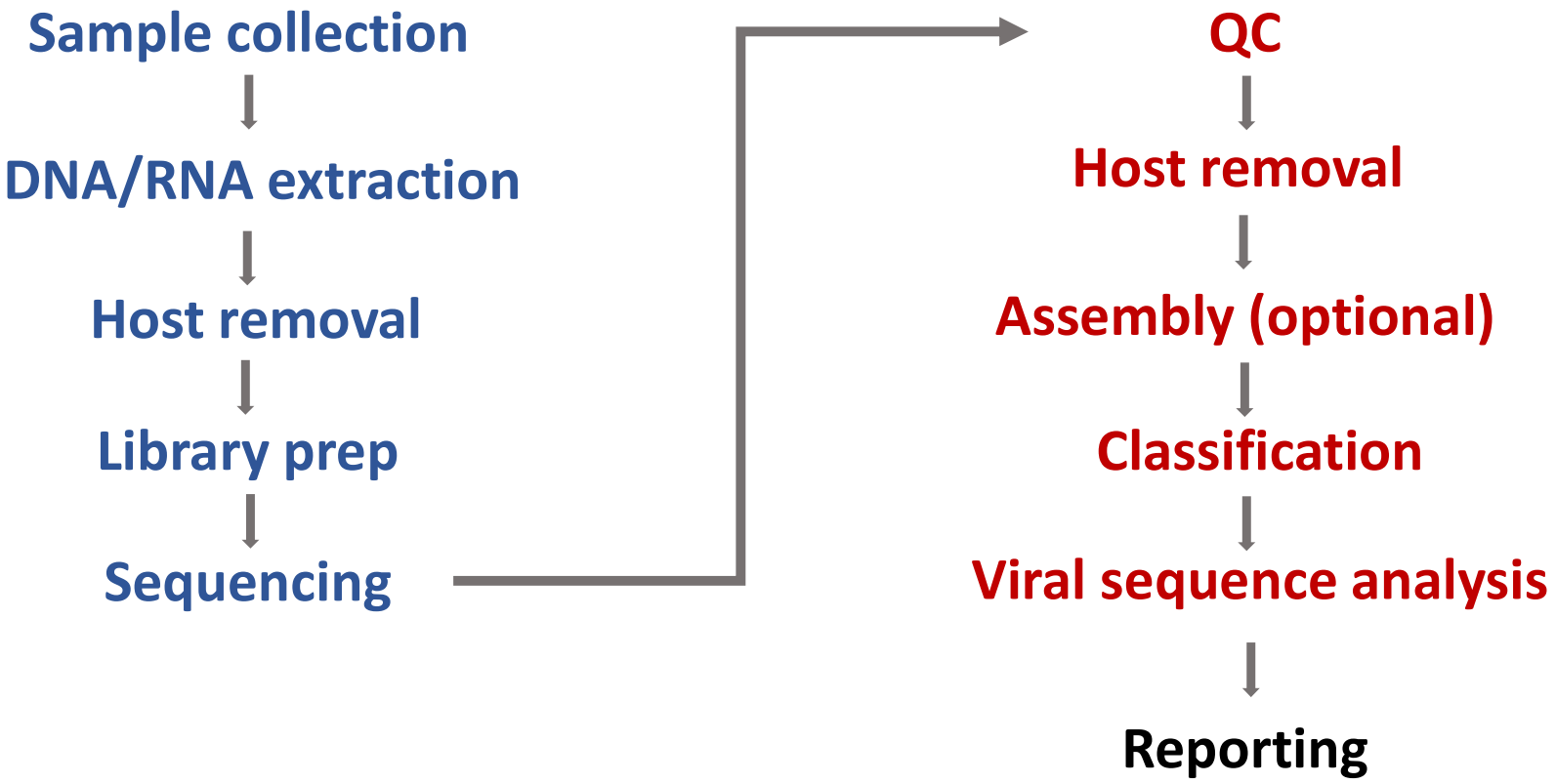
name	taxID	taxRank	genomeSize	numReads		numUniqueReads	abundance
SARS coronavirus	227859	leaf	29751	2	2	0.0	
Cyanophage S-RIM50	687803	species	174307	1	0	0.0	
Bat coronavirus BM48-31/BGR/2008	864596	species	29276	2	2	0.0	
Cyanophage S-RIM32	1278479	species	194437	1	0	0.0	
Tokyovirus A1	1826170	species	372707	1	1	0.0	

SUMMARY

Blast/aga assignment	Sequences count	Percentage	Legend
Not assigned	233	94.7%	
Severe acute respiratory syndrome-related coronavirus	4	1.63%	
Phytophthora parasitica virus	2	0.813%	
Kinglevirus lutadaptatum	1	0.407%	
Aphis citricidus bunyavirus	1	0.407%	
Mycobacterium phage Ariel	1	0.407%	
Brussowvirus 20617	1	0.407%	
Immutovirus immuto	1	0.407%	
Parvovirus NIH-CQV	1	0.407%	
Rauchvirus BPP1	1	0.407%	
Total	246	100%	

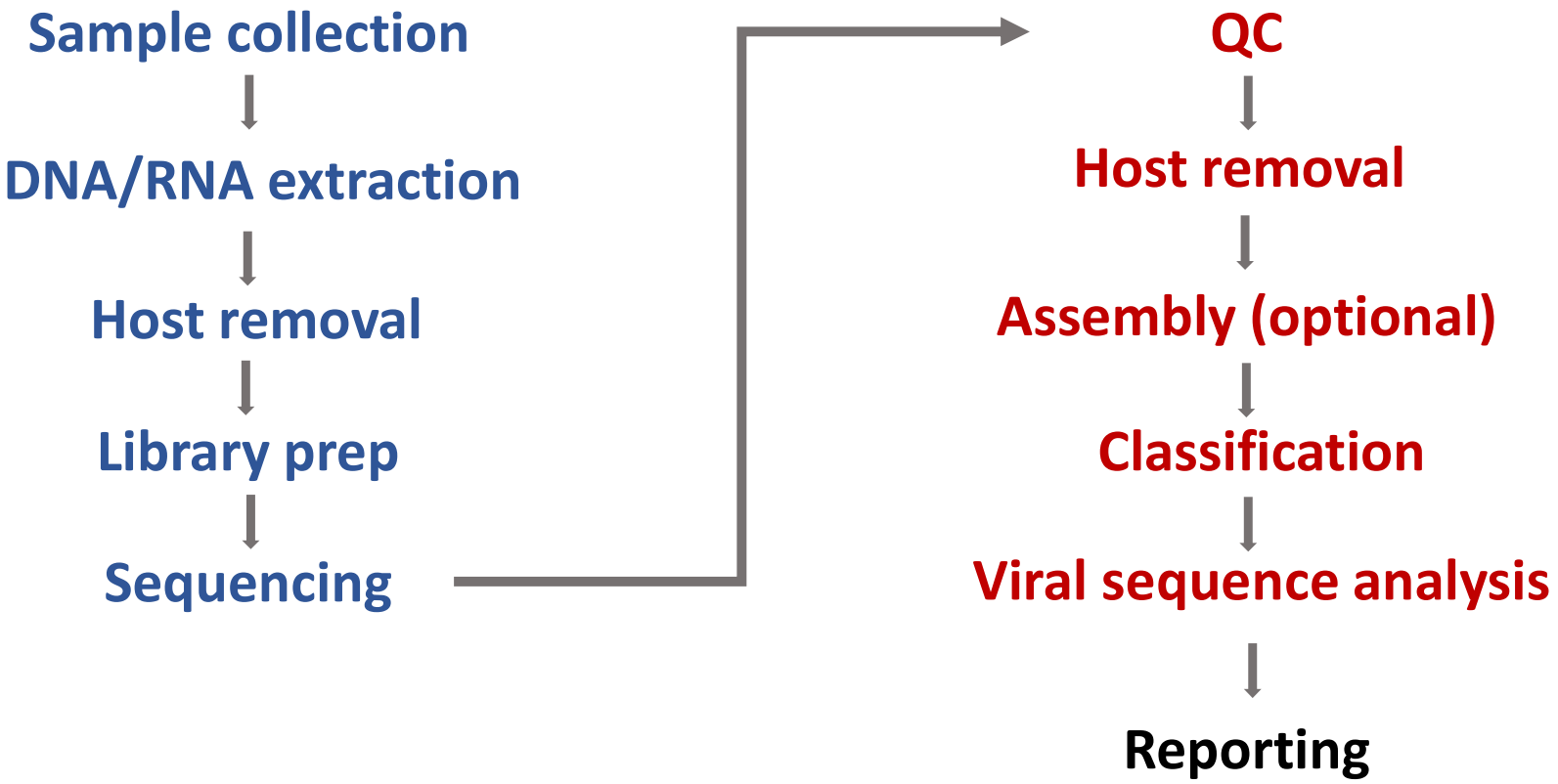
Protocol

```
metagen-fastqc.sh -t 8 \
-f SRR10971381_1.fastq.gz -r SRR10971381_2.fastq.gz \
-c hg38.fa
```



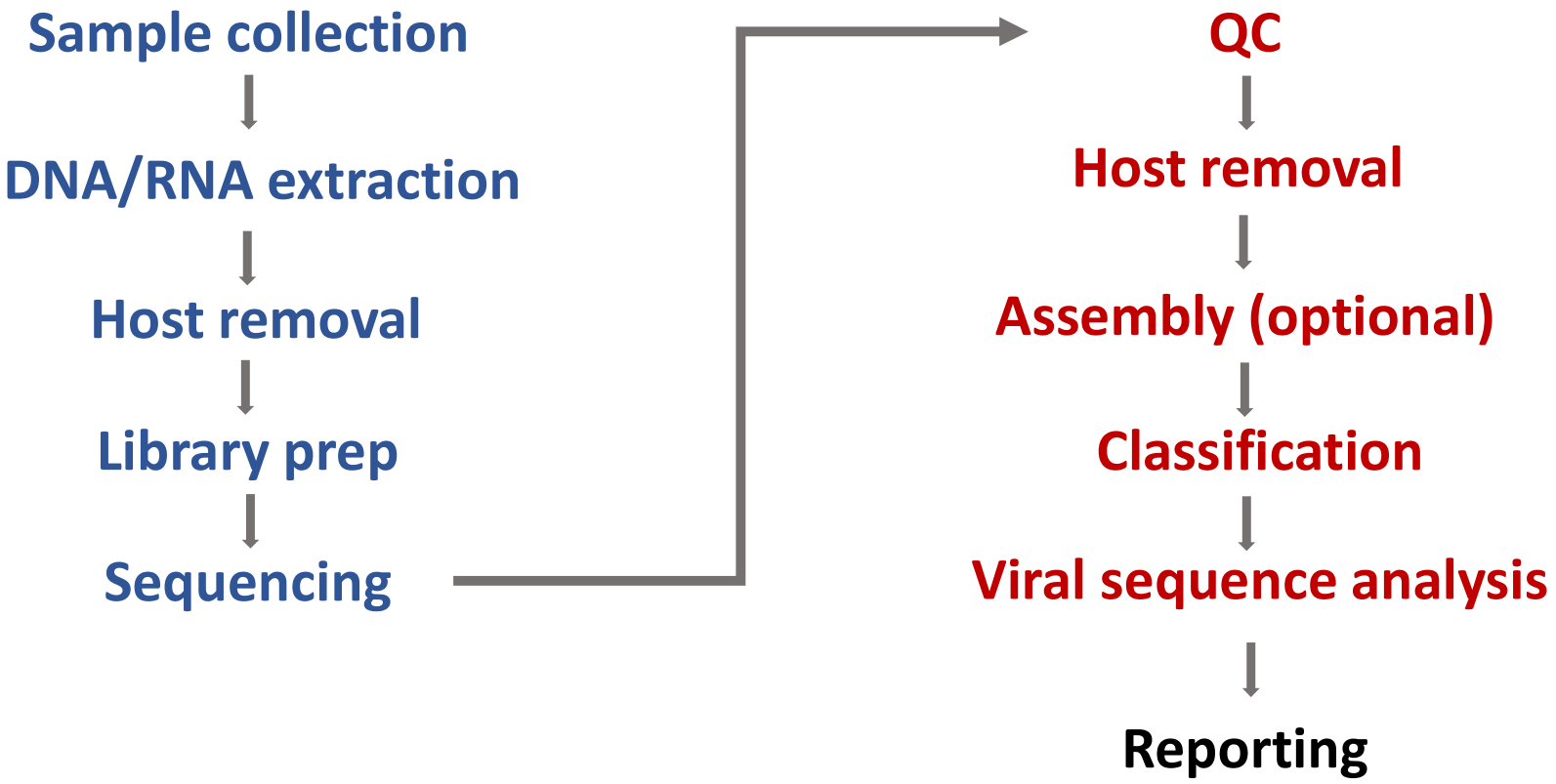
Protocol

```
spades.py --meta -k 21,33,55,63 \
-t 2 -m 8 \
-1 ~/Cristina/Metagenomics/Reads/SRR10971381_clean_sub_1.fastq.gz \
-2 ~/Cristina/Metagenomics/Reads/SRR10971381_clean_sub_2.fastq.gz \
–o ~/Cristina/Metagenomics/Assemblies
```



Protocol

```
centrifuge -p2 -f \
-x ~/Cristina/Metagenomics/Centrifuge_Db/Centrifuge-viral_db \
metaspades-raw.contigs.filtered.fasta \
-S contigs.filtered.viral-refseq.centrifuge
```



Choosing bioinformatics protocols for metagenomics

The protocol shown in the practical is probably 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!