Current challenges and best-practice protocols for microbiome analysis

Phuc Loi Luu, PhD

luu.p.loi@googlemail.com

July 07 2024

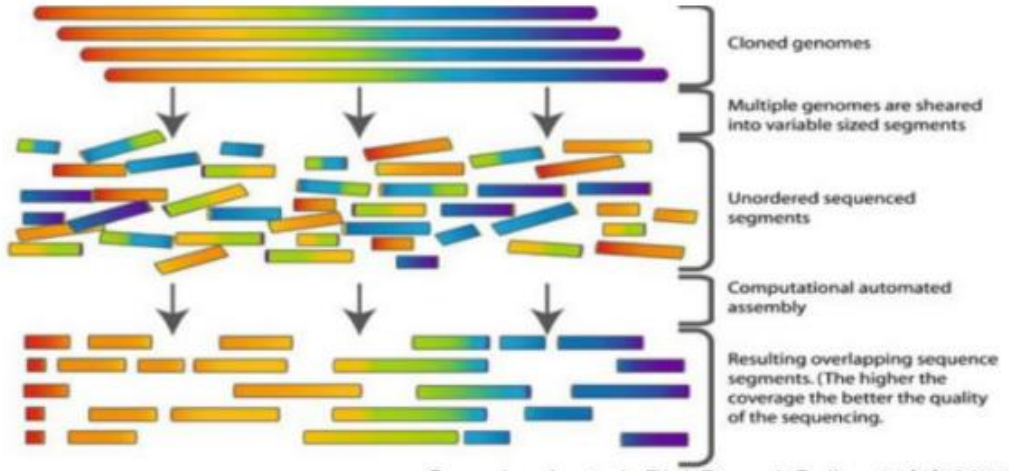
Content

- NGS-based microbial genotyping
- Gene amplicon/target/marker Metagenomics
- Shotgun Metagenomics
- Experimental challenges and solutions
- Sequencing and computational challenges
- Discussion

https://doi.org/10.1093/bib/bbz155

Next-Generation Sequencing (NGS):

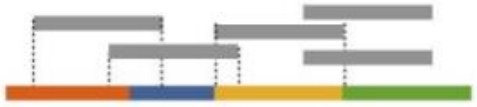
1) Reference 2) Massive Parallel Sequencing



Commins J. et al, Biol Proced Online 11(1) 2015

Mapping to reference sequence

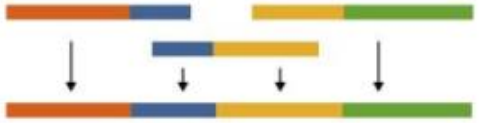
Recreate the genome with using prior knowledge as reference



Mapping is as good as reference used

De Novo assembly

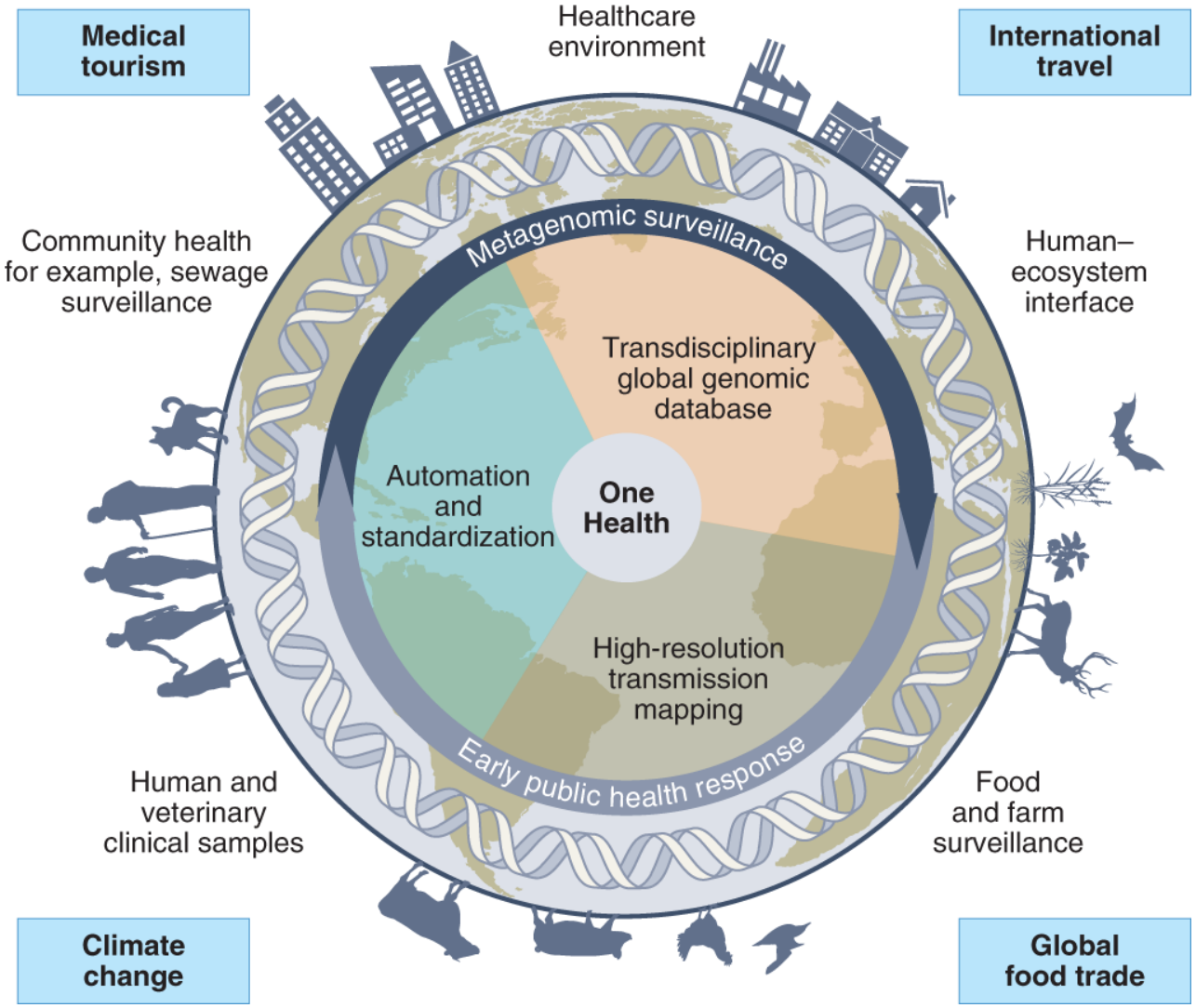
Recreate the genome with no prior knowledge



Problem with repeated regions, high coverage and long

https://pharmafeatures.com/cracking-genetic-ciphers-secondary-ngs-analysis/

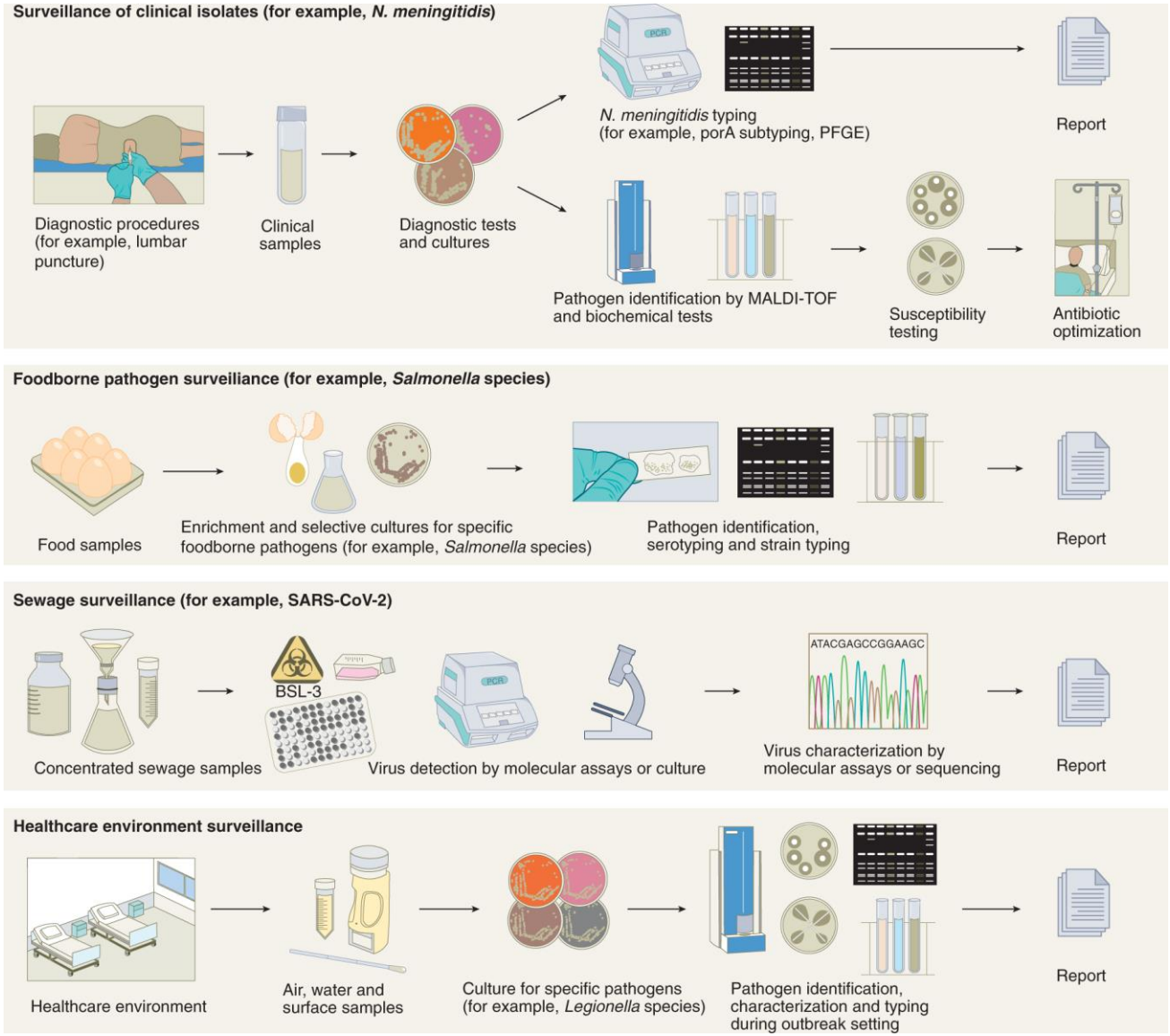
The opportunities for transdisciplinary One Health surveillance, utilizing metagenomic sequencing



A way to unify workflows and harmonize data. Key drivers for this integration are highlighted, including climate change and its impact on zoonotic reservoirs, globalization and international travel, medical tourism and global food trade.

https://www.nature.com/articles/s41564-022-01089-w

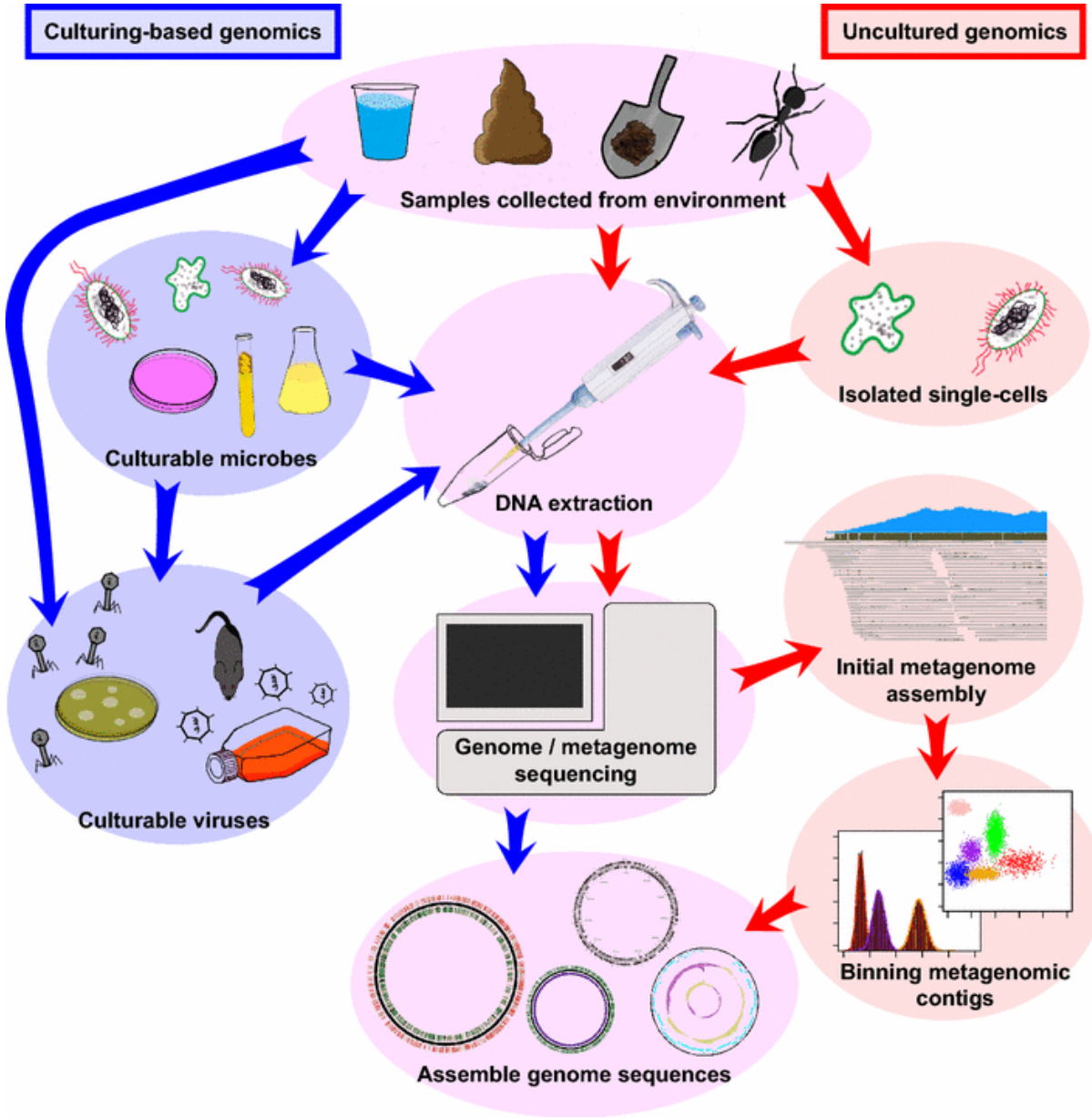
Culture-dependent/gold standard approach



The culturing step is followed by various typing and phenotypic assays

https://www.nature.com/articles/s41564-022-01089-w

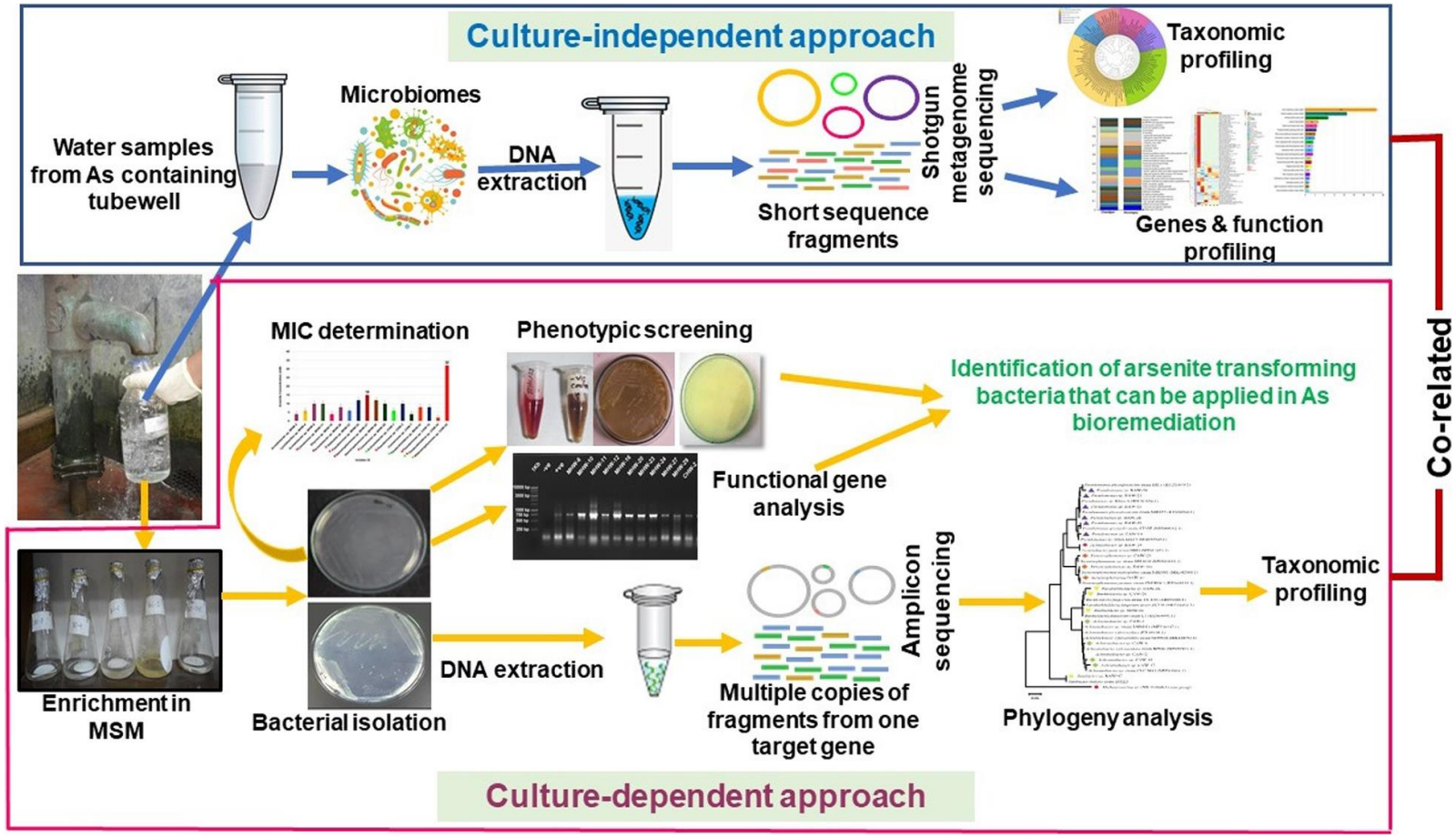
Culture vs unculture genomics



There are many variations of each protocol and additional steps, such as filtering samples according to molecular size cutoffs and normalization of data which are not illustrated in this diagram. The purpose is to illustrate simplified general steps to obtain uncultured genomes, which are common in most of the studies.

https://link.springer.com/article/10.1007/s00018-015-2004-1

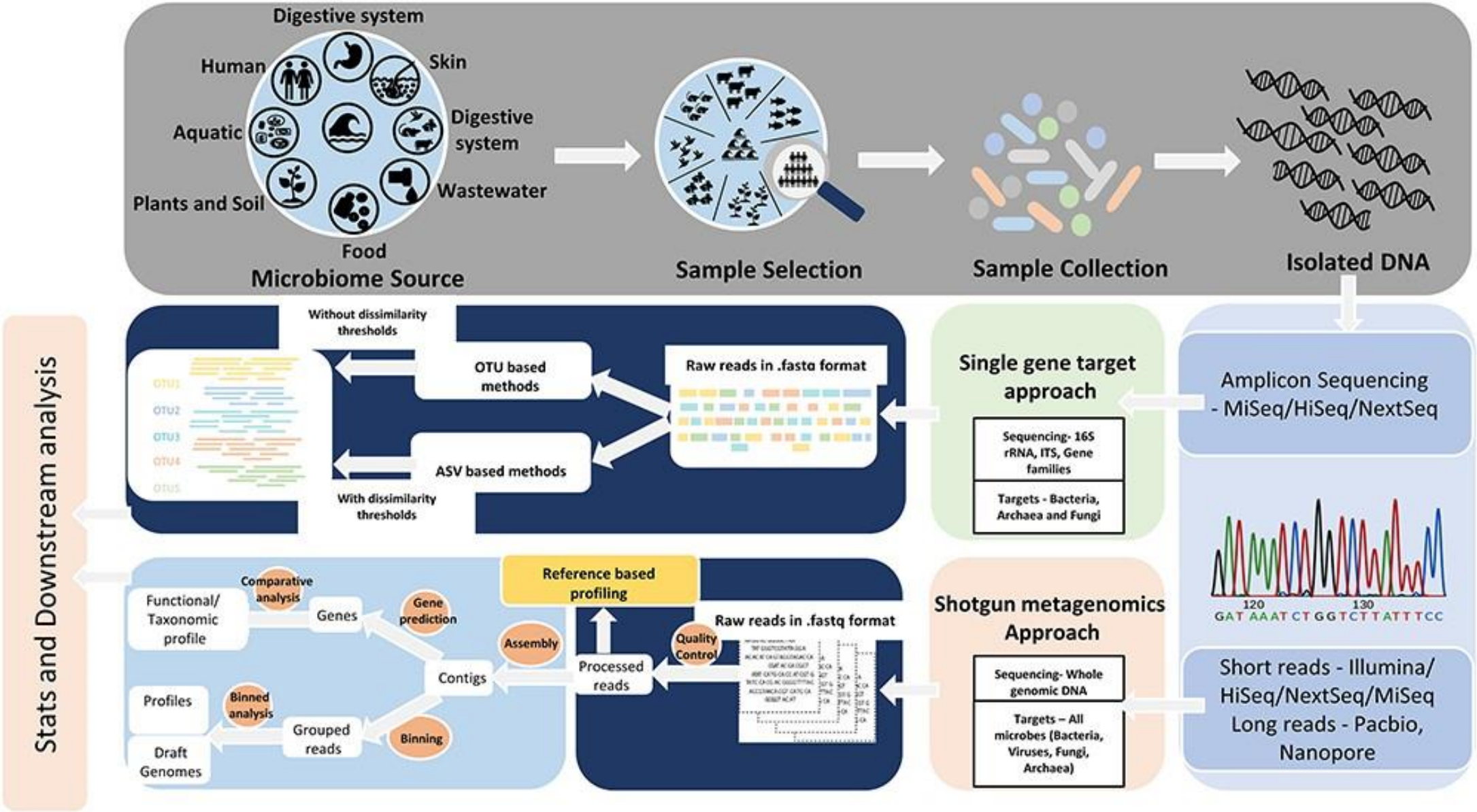
Metagenomic and culture-dependent



This study was carried out to assess the diversity and transformation potentials of the arsenic-affected groundwater microbiomes using both culture-dependent and independent (shotgun metagenomics) approaches

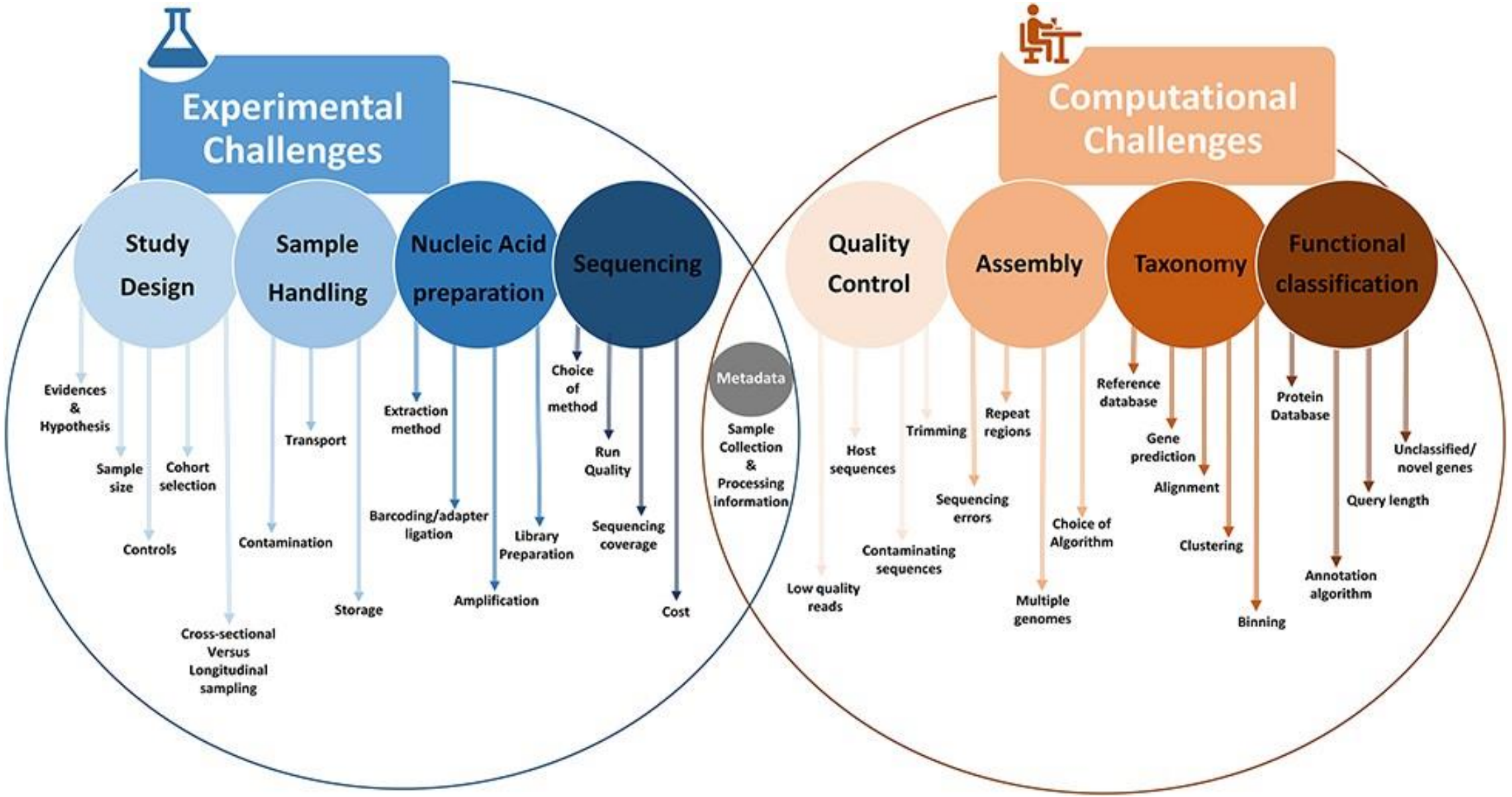
https://bmcmicrobiol.biomedcentral.com/articles/10.1186/s12866-023-02980-0

Targeted amplicon and metagenomic sequencing approaches



https://doi.org/10.1093/bib/bbz155

Overview of common experimental and computational challenges



Study design/experimental design

eliminating confounding effects

1. **Number of samples**: choosing appropriate sample sizes based on statistical principles can certainly help to avoid biases and spurious interpretations.
2. **Controls**: whether a signal is real and not just a stochastic
3. **Cross-sectional and longitudinal studies**
4. **Metadata**

Supplementary Table S1: Sample Metadata file

1. Sample ID - Sequencing Facility	
Sequencing ID	RB7486
Field Name	RB7486/TUM Campus Straubing
Description	Sample submitted by the TUM Campus Straubing, Germany
Privacy Risks	No
GenBank Structured Comment Synonym	GYA*
Data Categories	Sample shipment
Syntax	Alphanumeric
Data Source	HPIND
Comments	This data would be embedded in the GenBank record as a dbxref for linkage of HSC metadata records with GenBank sequences
2. Nucleic Acid Extraction Method	
Sequencing Assay Field ID	GA4
Field Name	Nucleic Acid Extraction Method
Description	Experimental protocol used to isolate nucleic acid fraction from the submitted sample for sequencing reaction.
Privacy Risks	No
Data Categories	Sequencing Sample Preparation
Example Values	Illumina standard method; CTAB/chloroform

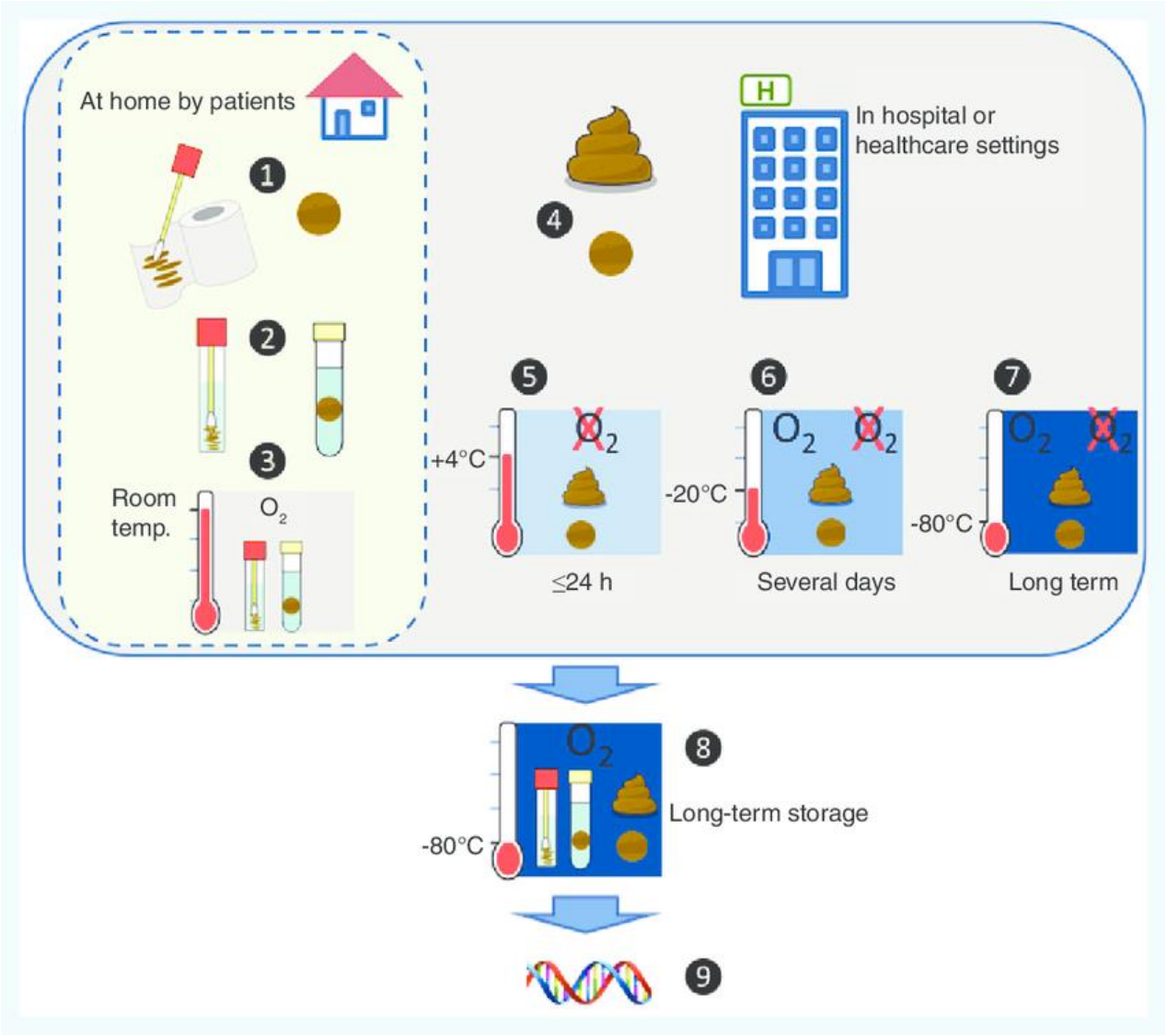
Sample collection and handling

a significant confounding factor

1. **Contamination**: time, temperature, humidity and proximity distance
2. **Transportation:** several chemical preservation methods and immediate freezing
3. **Storage and safety:** −80°C conserves microbiota diversity

Sample collection and handling

a significant confounding factor



Overview of methods that can be used for sample collection and storage before DNA extraction. Collections can be performed at home by patients, or in healthcare settings. When performed at home by patients (left) it is easier to collect stool samples rather than the complete bowel movement. If 16S analysis is to be performed, swabbing toilet paper can collect sufficient amount of material, whereas whole genome shotgun sequencing will require more material (~1 g) (1). In both cases, collected material will have to be immersed in stabilization buffer (2) before shipping to the lab that can then be done at ambient temperature and ambient atmosphere (3). In hospital or healthcare settings, the full specimen or samples can be collected (4). In both cases intermediary storage can be performed: (5) at 4°C in anaerobic atmosphere for less than 24 h, (6) at -20°C for several days (anaerobic or ambient atmosphere), (7) directly at -80°C for several months/ years (anaerobic or ambient atmosphere). Long-term storage of samples immersed in stabilization buffer, of the complete specimen or of the stool sample can be performed at -80°C in ambient atmosphere (8) before DNA extraction (9).

https://doi.org/10.2217/fmb.15.87

Nucleic acid extraction

effectively capture all types of microbes

There are two major extraction methodologies:

(i) mechanical lysis/bead beating

(ii) chemical lysis

Nucleic acid preparation

effectively capture all types of microbes

Widely used DNA isolation kits for library prep of short-read NGS:

1. **Nextera DNA Flex**
   - both large and small genome sizes with input DNA amounts of 100–500 ng and 1–500 ng
   - 96 multiplexed metagenomic samples
2. **Nextera XT**
   - 1 ng of input DNA samples
   - 384 uniquely indexed samples can be pooled and sequenced together
3. **TruSeq DNA PCR-Free**: little amounts of input DNA (~1 ng)

Nucleic acid preparation

effectively capture all types of microbes

1. Third-generation sequencing platforms (PacBio and ONT) prove to be more efficient
   - longer read sizes
   - Species/strain-level resolution
   - absence of DNA amplification-based biases
2. Pacbio Sequel II
   - enzymatic lysis of DNA for the extraction of longer DNA fragments
   - DNA output from > 10 Gb
3. ONT MinION
   - portable (size of a USB stick)
   - provides the agility to sequence samples from extreme conditions

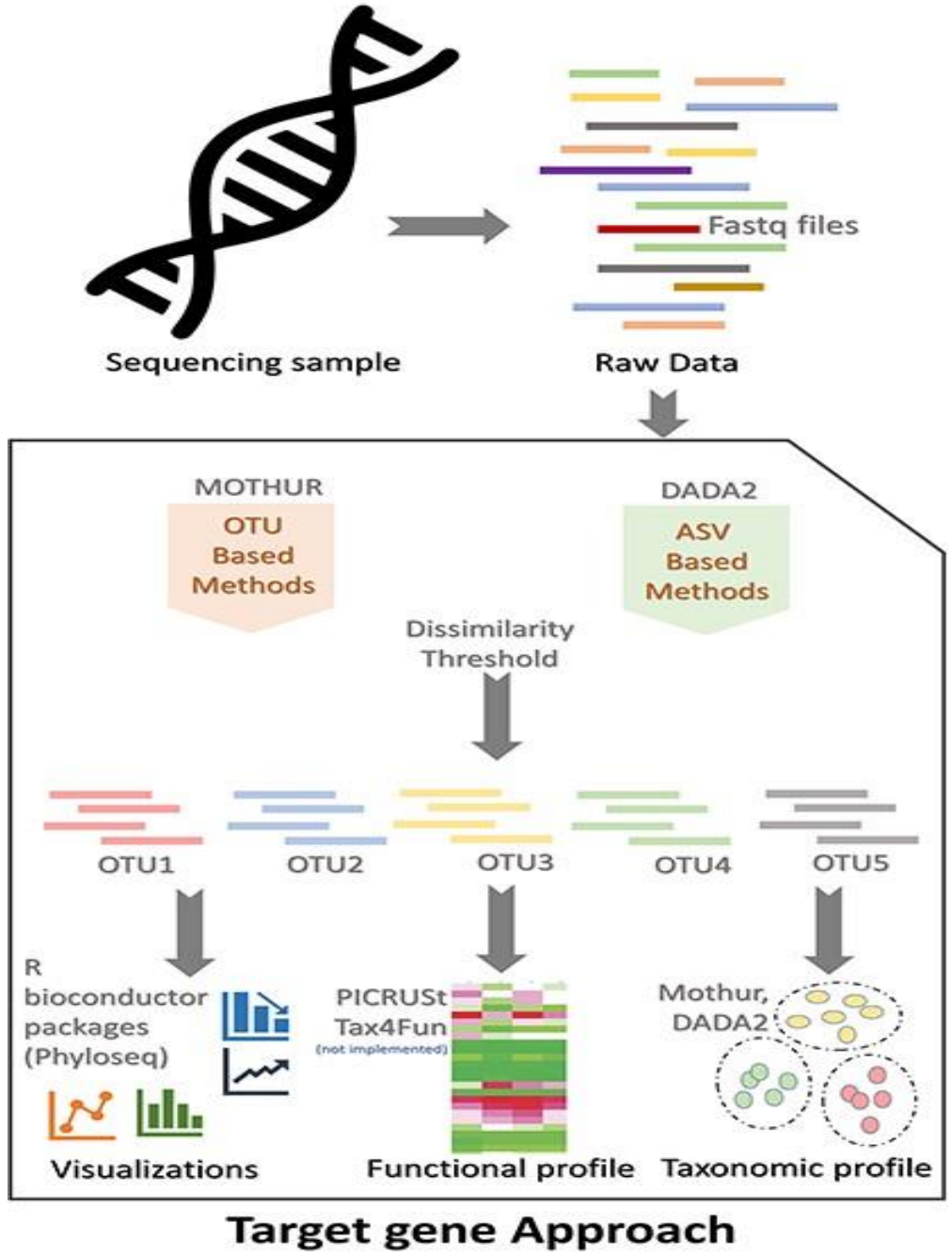
Challenges for amplicon sequencing analysis

- **Raw reads** are quality filtered and processed by either
  - OTU-based (**QIIME** and **Mothur**):
    - a predefined identity threshold (commonly 97%) into OTUs
    - lower taxonomic resolution
  - ASV-based (**DADA2**, **Deblur**, **MED**, and **UNOISE**) denoising methods utilizing a dissimilarity threshold

- **Taxonomy**: Microbial communities are identified through a rigorous protocol that results in multiple pangenome alignments using customized databases (**SILVA**, **Greengenes** and **RDP**)

- **Functional annotation**: **Tax4Fun** (R-based algorithm utilizing SILVA) or **PICRUSt** (clusters protein sequences based on KEGG or COG gene families and 16S rRNA gene copy numbers).

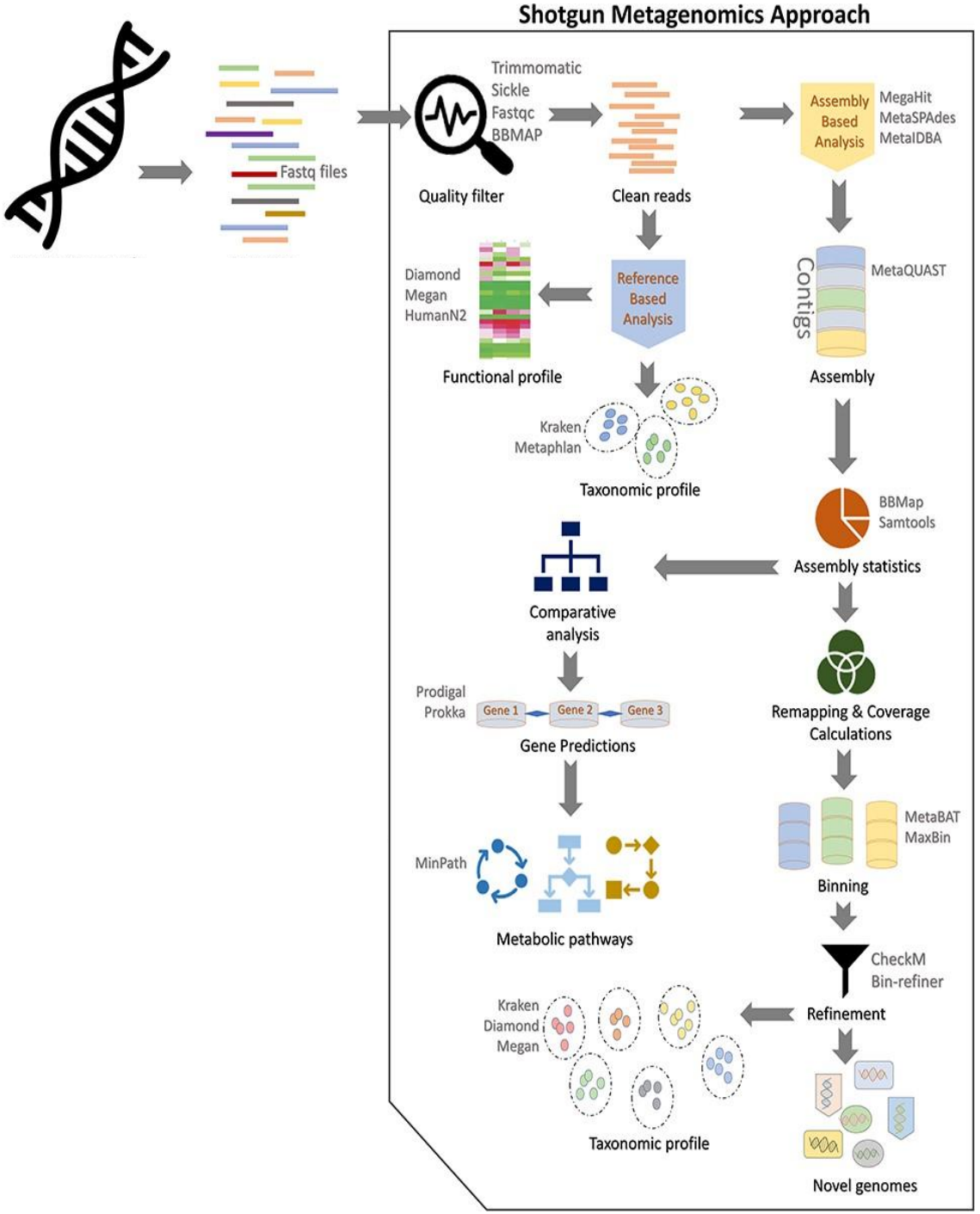
- Data visualization: R Bioconductor package **phyloseq**



Challenges in shot-gun metagenomics analysis

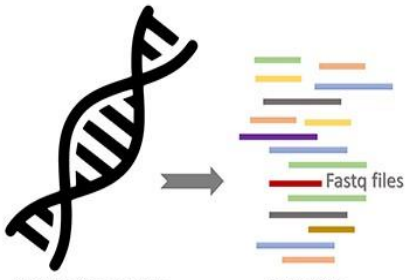
Quality filtering

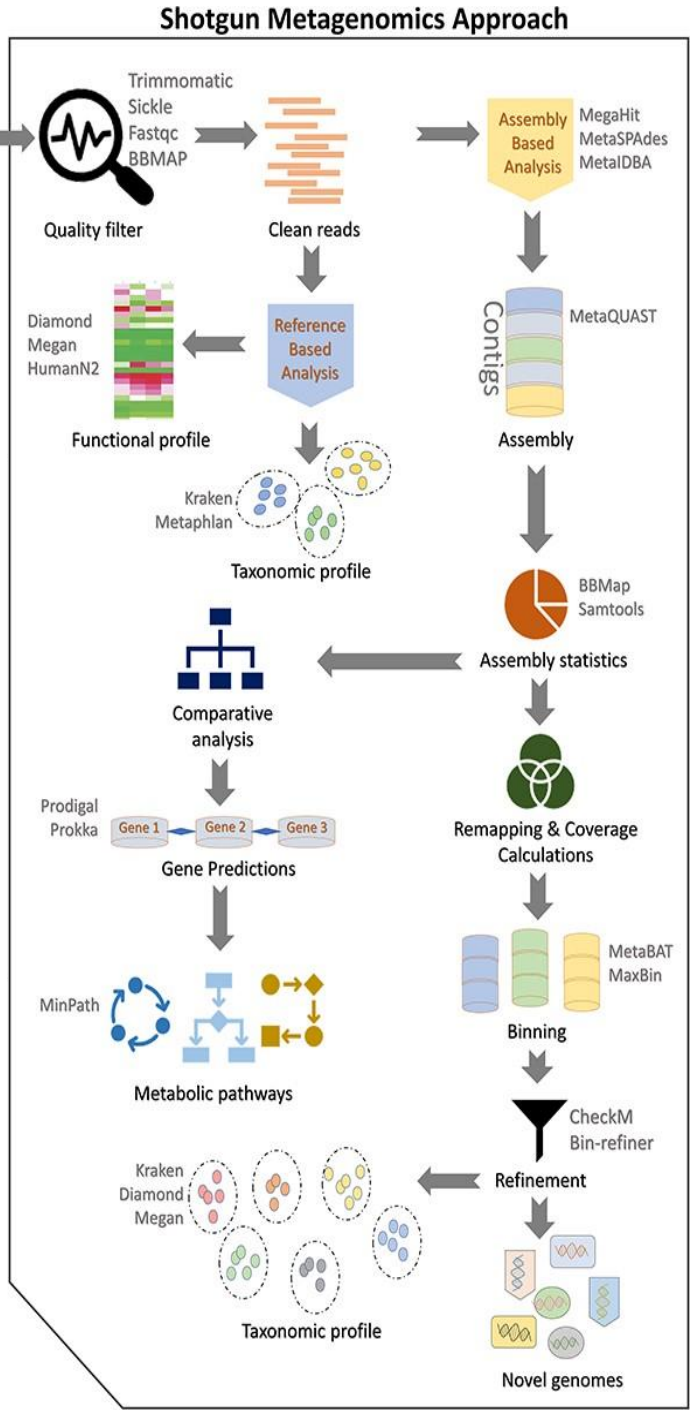
- **Tools for trimming**
  - Trimmomatic
  - Sickle
  - BBTools



Challenges in shot-gun metagenomics analysis

Reference-based analysis



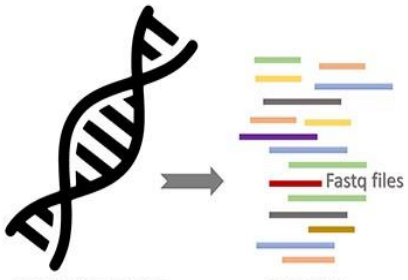


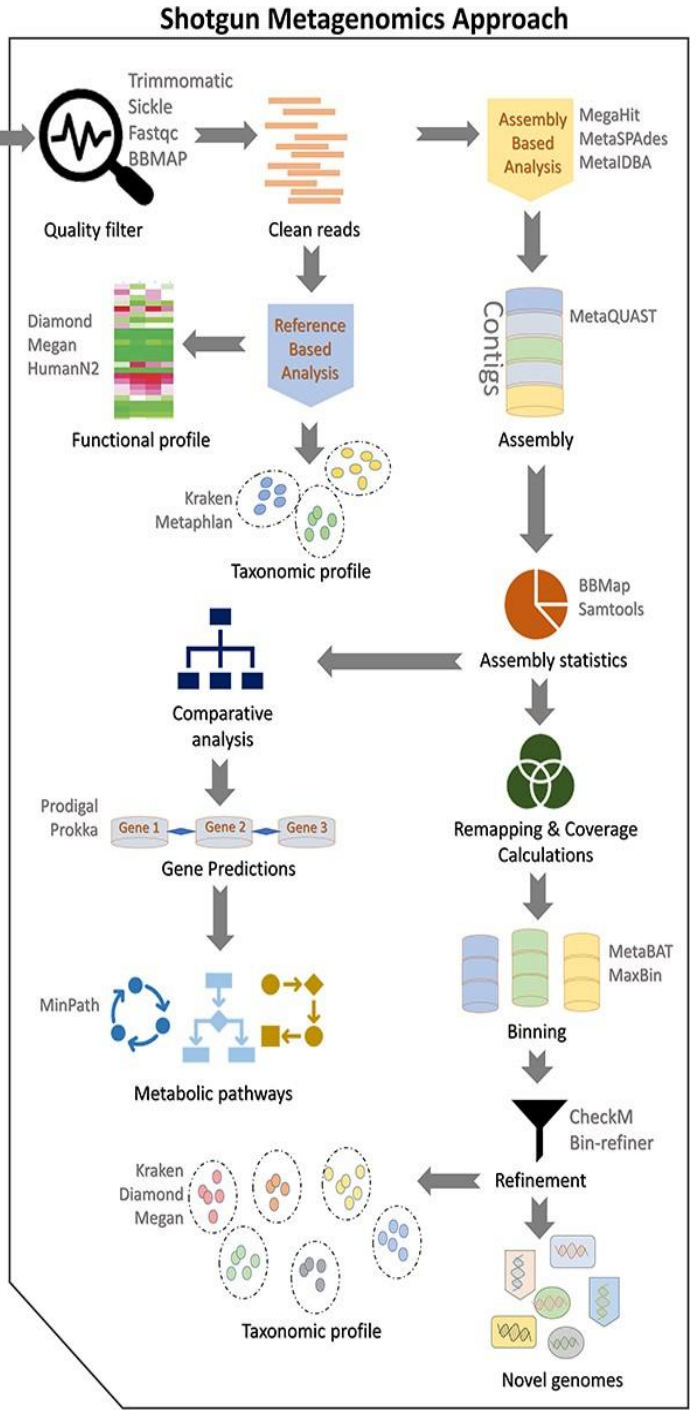
• **Taxonomy**: Compositional profiling of communities from metagenomic sequencing data can be optimally done by either using unique clade-specific marker genes identified from 3000 reference genomes (**MetaPhlAn**) or by exact alignments of *k*-mers alongside a classification algorithm (**Kraken**).

• **Functional annotation**: The functional profiling of metagenomic communities can be optimally performed using **HUMAnN2** or **Megan** pipelines. For long reads the **DIAMOND** sequence aligner can be used alone or with Megan to perform pairwise and frameshift alignments.

Challenges in shot-gun metagenomics analysis

Assembly-based analysis

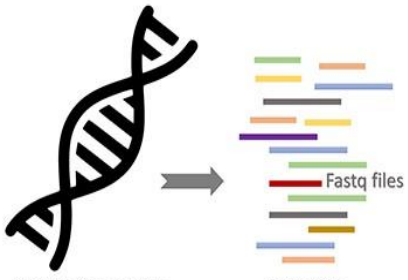


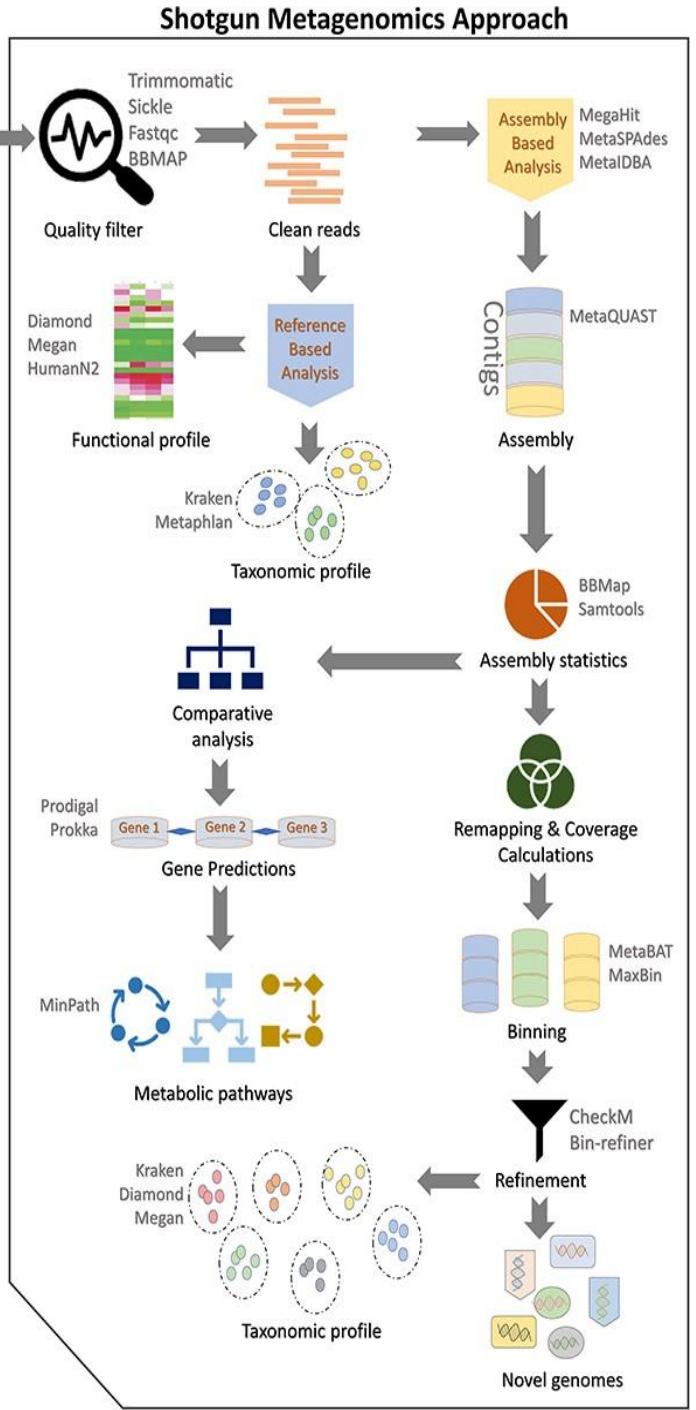


- The three most optimal assembling algorithms: **MegaHit**, **MetaSPAdes** and **MetaIDBA**
- **Contig assembly**: assembled reads are clustered into contigs and evaluated by **MetaQUAST**
- **Assembly statistics**: This step is a prerequisite of remapping/coverage calculations and comparative analysis with **BBMap** (both short- and long-read sequences from Illumina, PacBio, or MinION)
- **Comparative analysis**: incorporates algorithm-based gene predictions and metabolic pathway identifications. **Prokka** annotates the data by predicting genes using **Prodigal** and then performs functional annotation on these genes. The **MinPath** algorithm could be implemented for biological pathway reconstructions based on protein family predictions.

Challenges in shot-gun metagenomics analysis

Assembly-based analysis

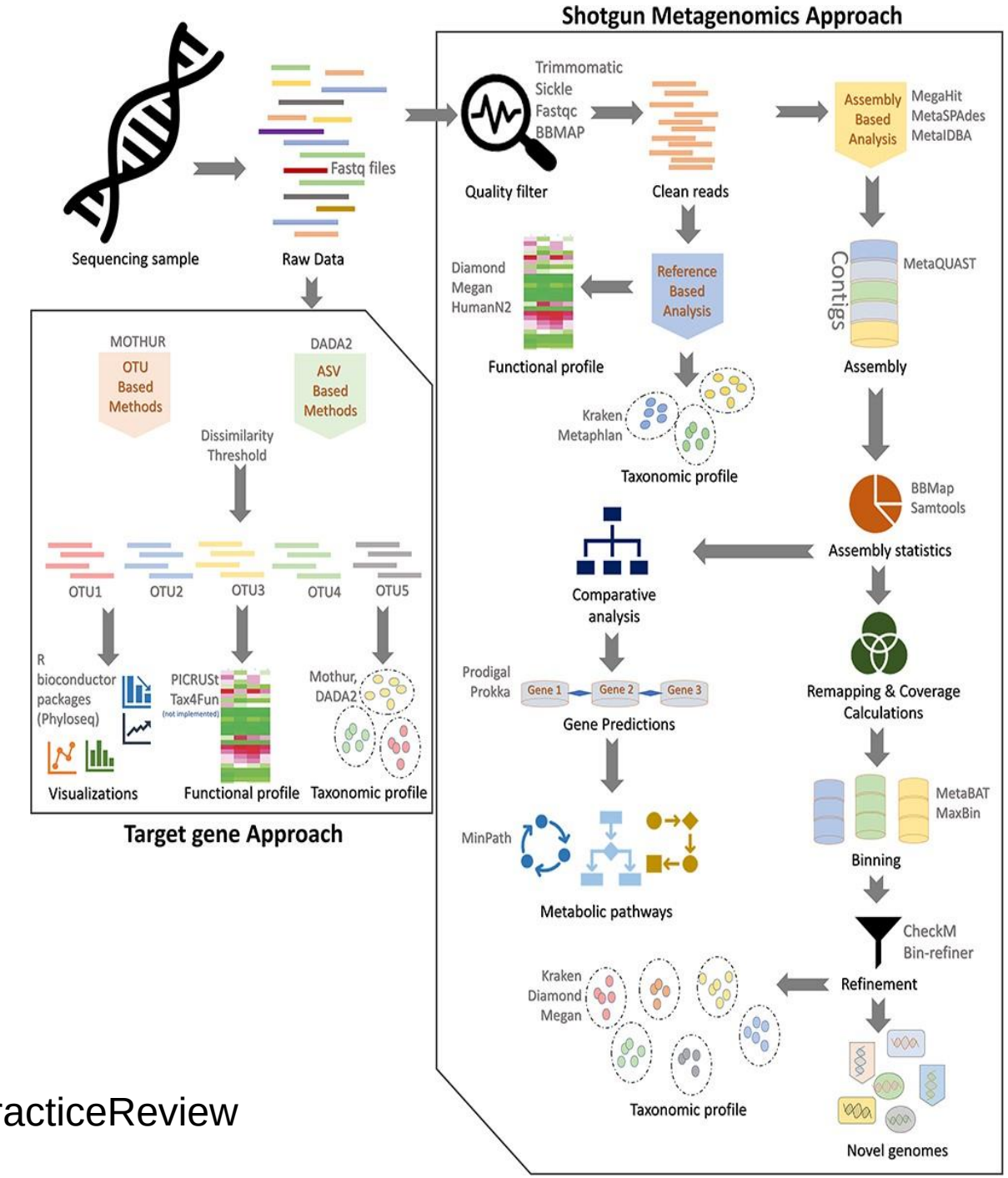




- **Binning**: binning or grouping of generated contigs is done before further downstream analysis. Either **MetaBAT** with an adaptive binning algorithm or **MaxBin** that utilizes an EM algorithm could be used for metagenomic contig binning.

- **Refinement**: generating taxonomic profiles and annotation of any novel genomes present in samples. Both **CheckM** and **bin-refiner** are optimally used for estimating genome completeness and contamination. Taxonomic profiles and novel genome identification can be optimally performed using **Kraken** and Diamond algorithms with or without the **Megan** pipeline.

Thank you!



https://github.com/grimmlab/MicrobiomeBestPracticeReview