Data Analysis Working Group

Workshop Series FA24

Meeting 1

09/27/24

OUR TEAM

Polina Tikhonova
Director

Human-microbial Interactions

Bioinformatics

Emily R. Davenport

Susan Tian
Secretary

Gut Microbiome

BMMB

Jordan Bisanz

Jamie Spychalla
Social chair

Disease Ecology

Plant Pathology

Sharifa Crandall

Daniela Betancurt
Workshop chair

Microbial Ecology

BMMB

Jordan Bisanz

OUR EVENTS 2024-2025

WORKSHOPS FA24

AVBS 106

Sept, 27 (TODAY)	*Introduction to Metagenomics analysis, Setting up an environment*
Oct, 25	*Metagenomics data processing*
Nov, 22	*Metagenomics Differential Abundance Analysis*

WORKSHOPS SP25

TBD	*Functional analysis,etc*

CODING CLINICS

- *A co-working space, where everyone can come and focus on their project*
- *Receive help with troubleshooting*
- *Ask for an advice for your analysis*

Oct, 4th 9:30-10:45 AM	*AVBS 106*
TBD	

“Give DAWG a bone” AWARD!

Fund your project **& Analyze** with DAWG!

HOW MUCH

Up to $5000 for sequencing your data: microbiome, metagenome, mycobiome or other microbes!

WHO

Postdoctoral scholars, graduate students, undergraduate student affiliated with One Health Microbiome Center!

WHEN

Each spring. Announcement is posted in the newsletter.

KEEP IN TOUCH!

@DAWGPSU

PSU-DAWG@LISTS.PSU.EDU

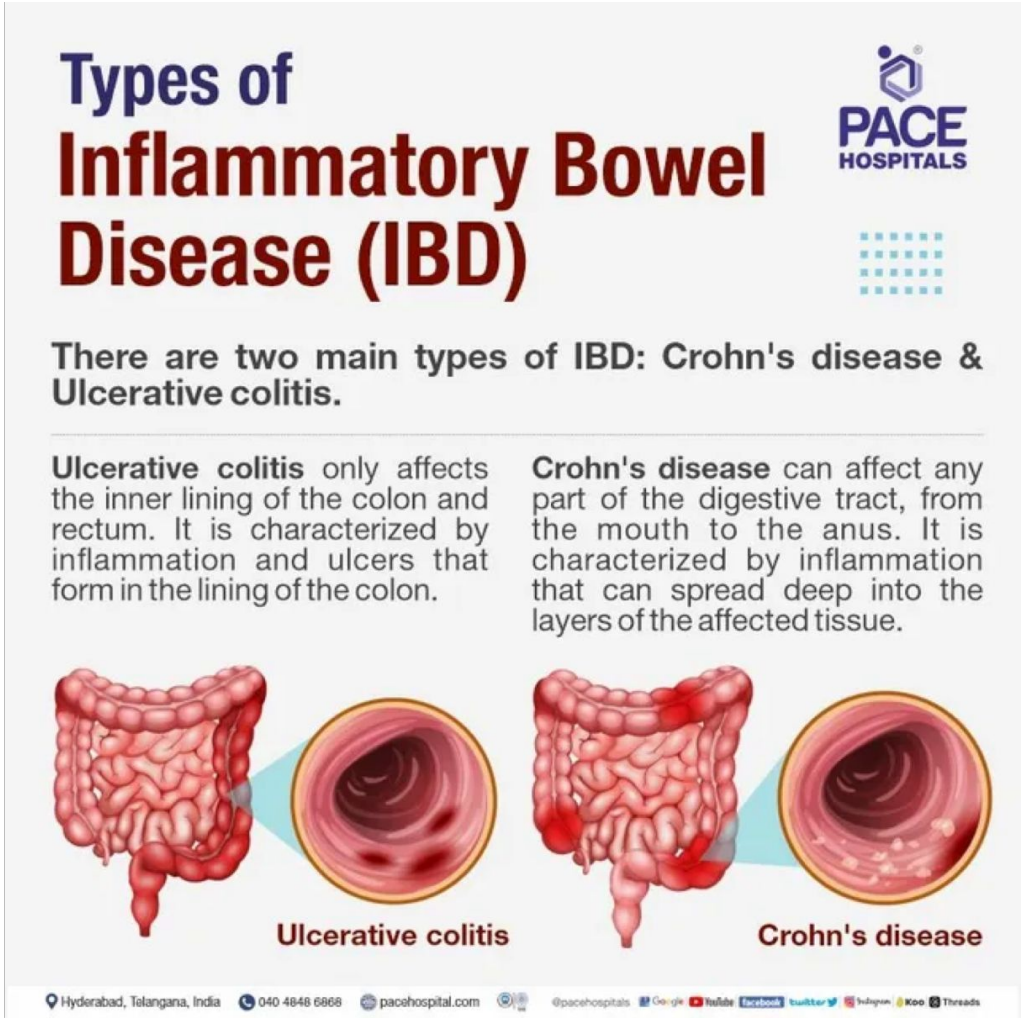
Please, fill our pre-workshop questionnaire to help us to adjust your needs! **https://bit.ly/DAWG_q1**

Microbiome and metabolome features in inflammatory bowel disease via multi-omics integration analyses across cohorts

DAWG Workshop Sep 27 2024

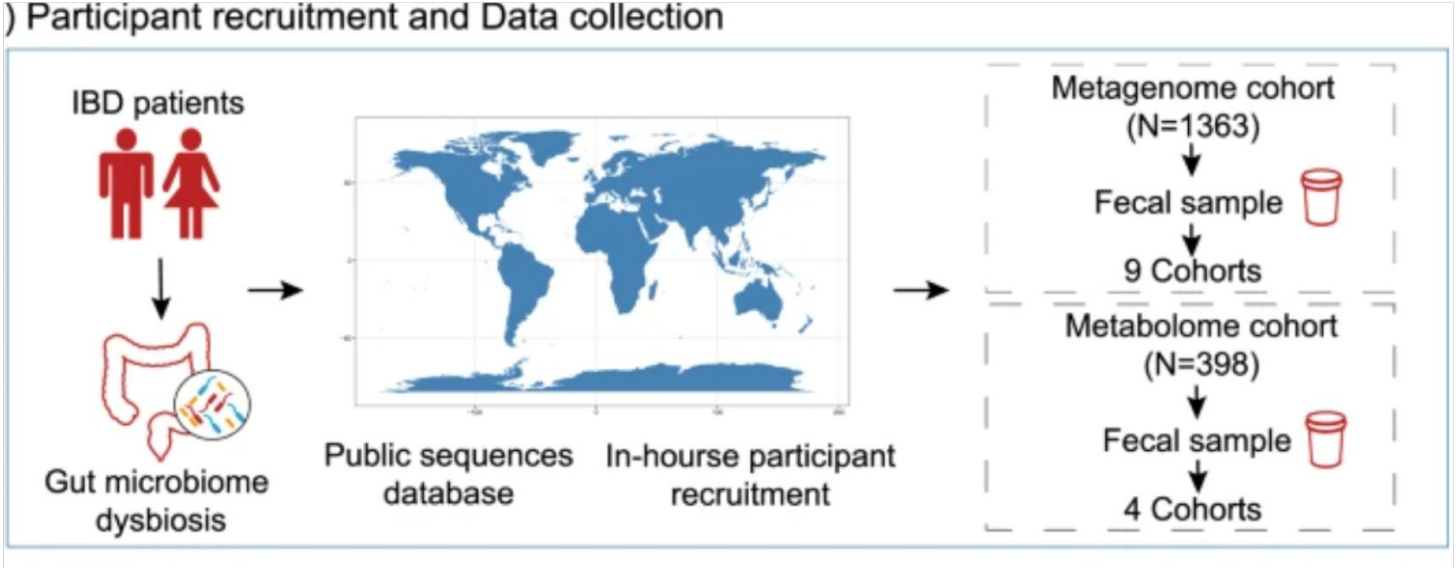
IBD

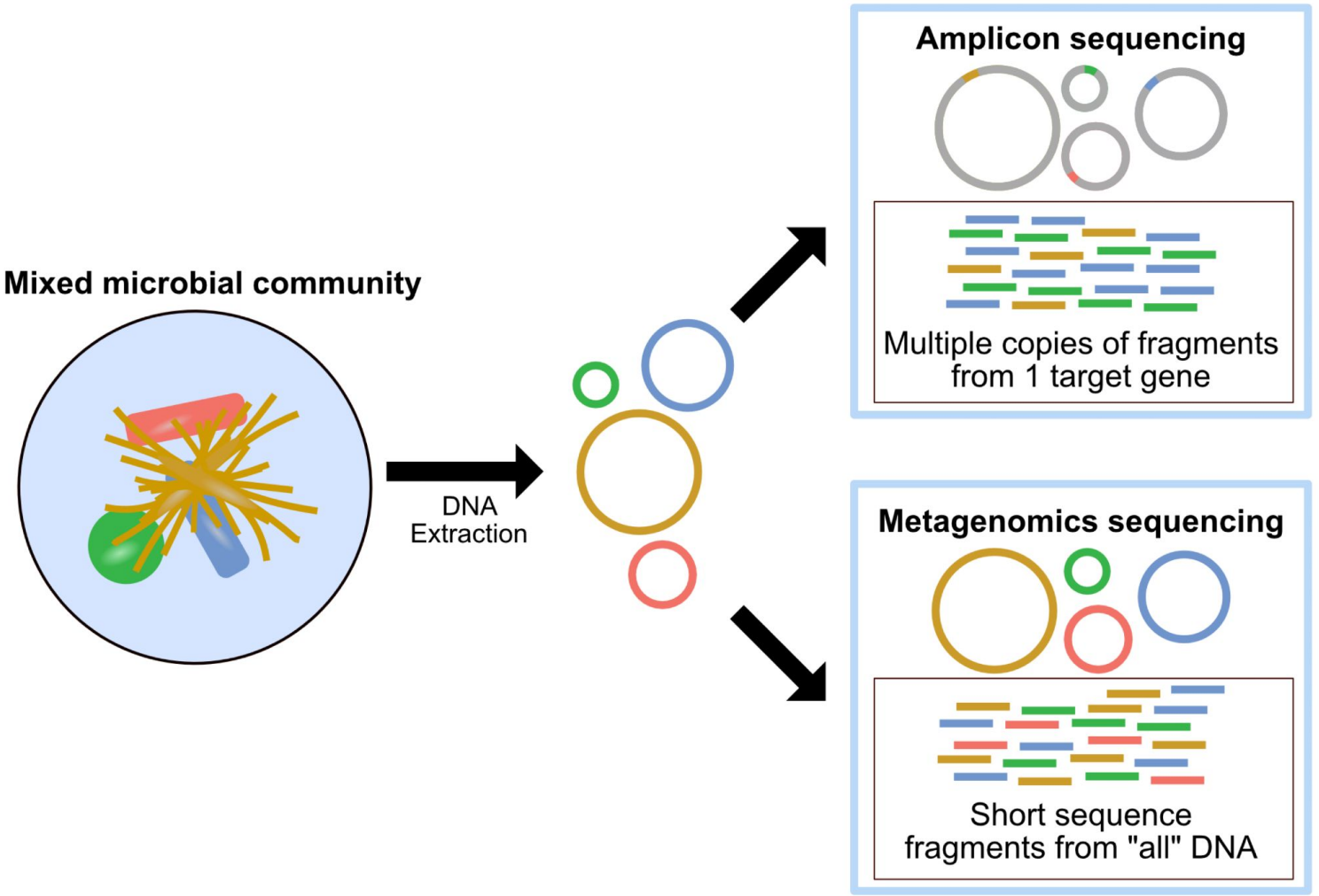
- Inflammatory bowel disease: chronic inflammatory conditions that affects the GI tract and includes two main forms: Crohn’s disease(CD) and Ulcerative colitis(UC)

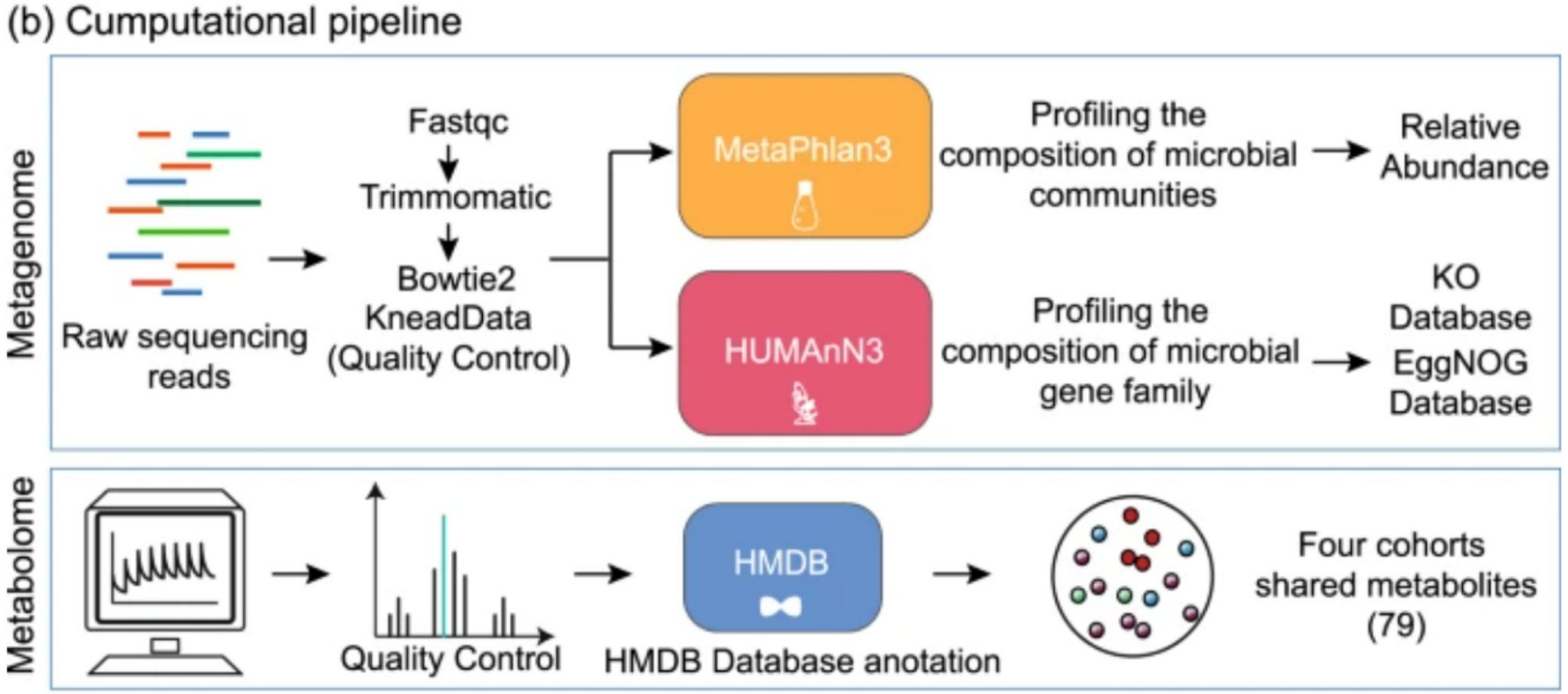


CCIA

- Goal: identify disease-associated species and metabolite as non-invasive biomarkers for IBD
- Approach: CCIA: cross-cohort integrative analysis to solve the challenge of variations in multiple datasets

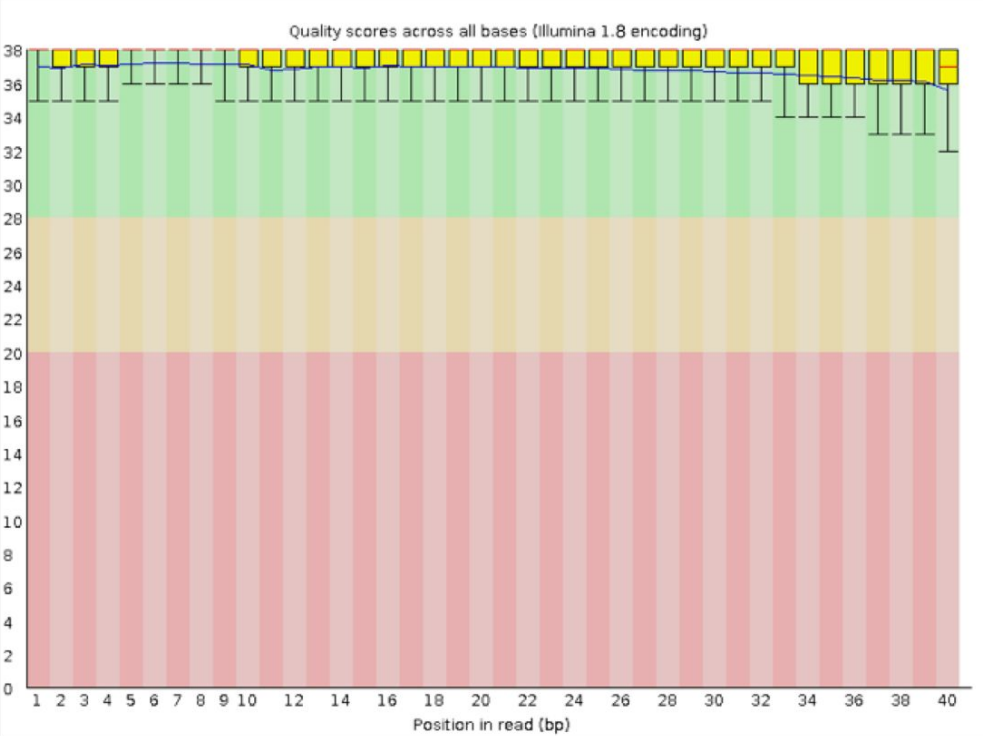



Mixed microbial community
DNA
Extraction
Amplicon sequencing
Multiple copies of fragments
from 1 target gene
Metagenomics sequencing
Short sequence
fragments from "all" DNA


(b) Cumputational pipeline
Metagenome
Raw sequencing reads
Fastqc
Trimmomatic
Bowtie2
KneadData
(Quality Control)
MetaPhlan3
HUMAnN3
Profiling the composition of microbial communities
Relative Abundance
Profiling the composition of microbial gene family
KO Database
EggNOG Database
Metabolome
Quality Control
HMDB
HMDB Database anotation
Four cohorts shared metabolites (79)

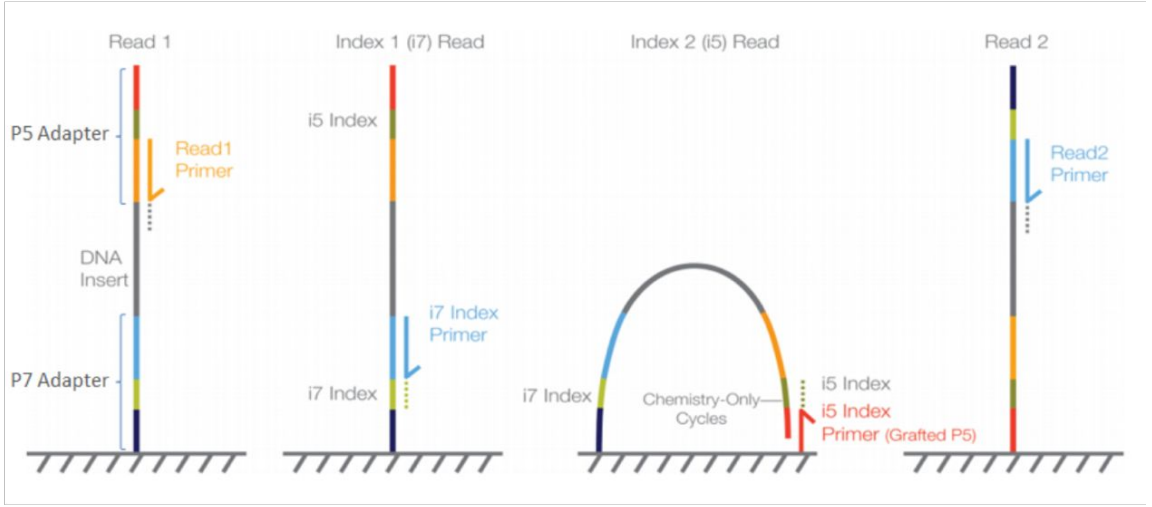
Data processing—FastQC

- FastQC provide a simple way to do quality control checks on raw sequencing data from high throughput sequencing pipelines
- FastQC's functions include:
  - import data (BAM, SAM, or FastQ files/ any variant)
  - a quick overview to locate the problem area
  - summary graph and tables for the data
  - export results to HTML based permanent repots



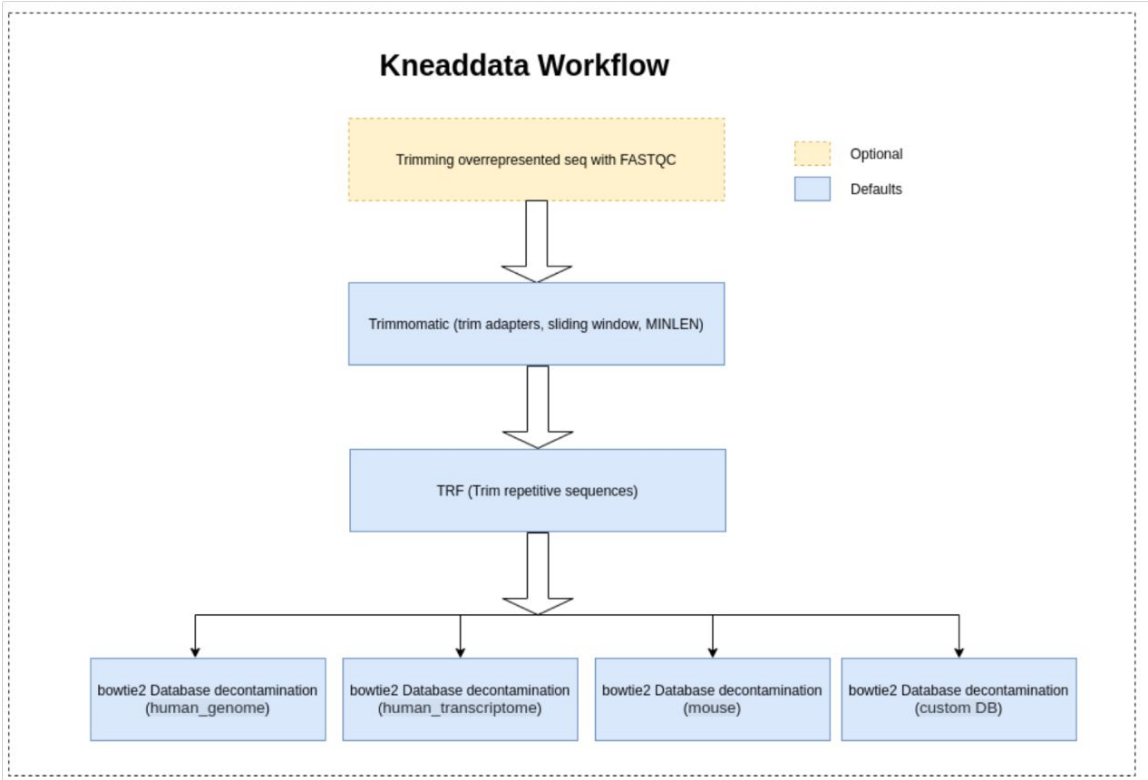
Data processing—Trimmomatic

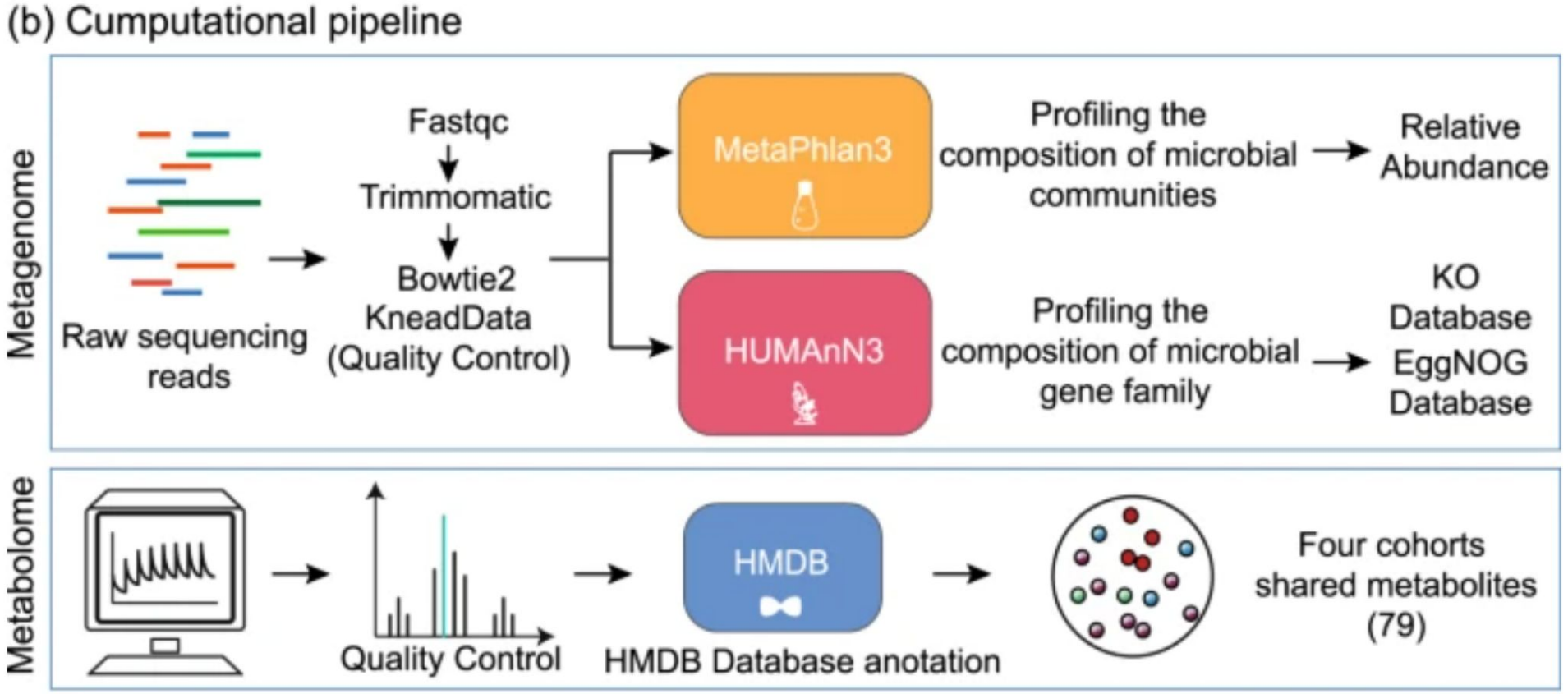
- Trimmomatic: a flexible read trimming tool for NGA data
- to perform quality trimming and adapter clipping—to prevent interfering the down stream analysis such as sequence alignment to the reference
- support both single end and paired end trimming



Data processing—Bowtie2 and KneadData

- Bowtie2 is an ultrafast and memory efficient tool for aligning sequence reads to long reference sequences
- KneadData: aim to in silica separation of bacterial reads from the “contaminant” reads (host VS bacteria)—bacteria only or human only

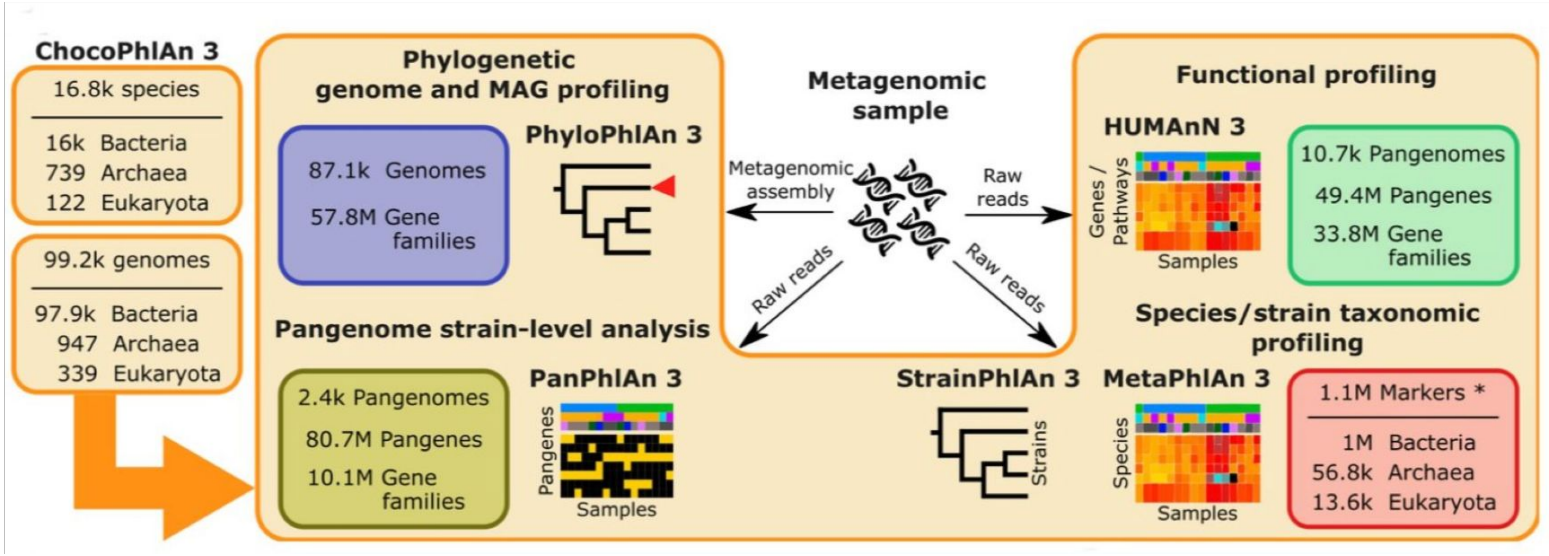



(b) Cumputational pipeline
Metagenome
Raw sequencing reads
Fastqc
Trimmomatic
Bowtie2
KneadData
(Quality Control)
MetaPhlan3
HUMAnN3
Profiling the composition of microbial communities
Relative Abundance
Profiling the composition of microbial gene family
KO Database
EggNOG Database
Metabolome
Quality Control
HMDB
HMDB Database anotation
Four cohorts shared metabolites (79)

MetaPhlAn: Metagenomic Phylogenic Analysis

- for species-level microbial profiling (bacteria, archaea, eukaryotes, and viruses) from **shotgun sequencing data**.

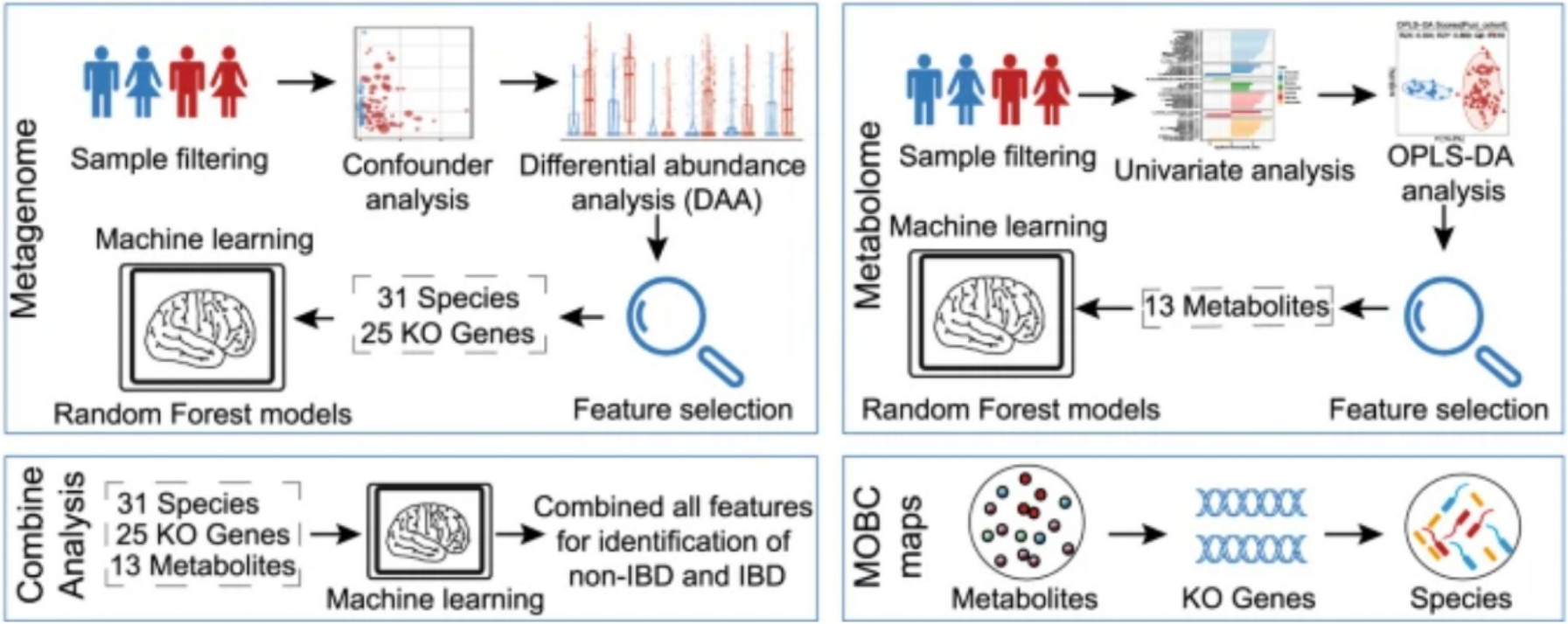
- uses **clade specific marker genes** identified from more than 1M microbial genomes. The most recent version MetaPhlAn4 also support the abundance estimation



HUMAnN: HMP Unified Metabolic Analysis Network

- profile the abundance of microbial metabolic pathway and other molecular functions from metagenomic or metatranscriptomic sequencing data
- =what are the microbes in my community-of-interest doing/capable of doing?
- workflow:
  - run MetaPhlAn and ChocoPhlAn pan genome database to get community functional profile stratified by known and unclassified organisms
  - **align the sequences against databases** of genomes and pathways such as UniRef (gene facility definitions), MetaCyc (pathway definitions by gene family), and MinPath (identify the set of minimum pathway) **using accelerated mapping tools** such as Bowtie2 (for nucleotide level searches) and Diamond (for translated/protein searches)

(c) Data Analysis



Installing miniconda

```
mkdir -p ~/miniconda3
curl https://repo.anaconda.com/miniconda/Miniconda3-latest-MacOSX-arm64.sh
-o ~/miniconda3/miniconda.sh
bash ~/miniconda3/miniconda.sh -b -u -p ~/miniconda3
rm ~/miniconda3/miniconda.sh
```

https://docs.anaconda.com/miniconda/

Initialize Miniconda following installation

~/miniconda3/bin/conda init bash

~/miniconda3/bin/conda init zsh

Creating and activating a conda environment

conda create --n <my-env>

conda activate <my-env>

conda env list (This allows you to see all the environments on your machine)

For example:

conda create – n test01 python= 3.4

conda activate test01

https://conda.io/projects/conda/en/latest/user-guide/tasks/manage-environments.html

Conda libraries

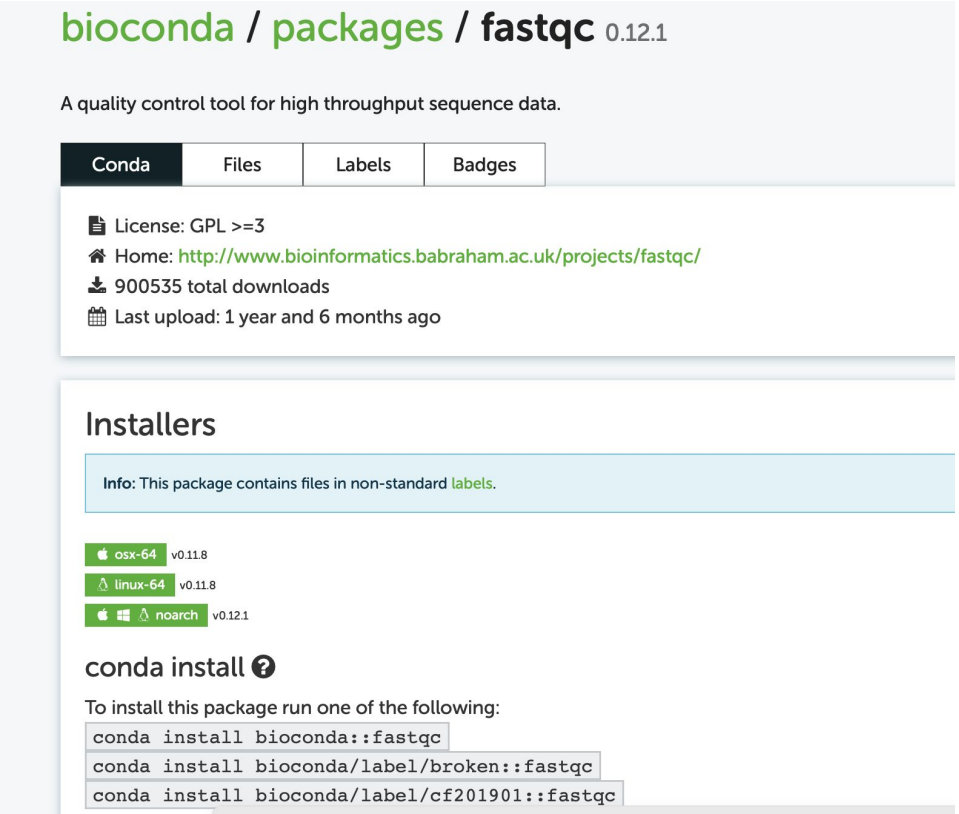
conda list

conda install <library name>

For example

conda install bioconda

conda install bioconda::trimmonmatic



You can also use the pip install command in the conda environment

Export and Recreate Environment with YAML File

conda env export > testenv.yml

conda env create -f environment.yml

https://saturncloud.io/blog/how-to-create-a-conda-environment-based-on-a-yaml-file-a-guide-for-data-scientists/