Comparison of Metagenomic Tools in Gut Microbiome Analysis of COVID-19 Patients

Muhamad Rifki Ramadhan

Supervisor: Dr. Janetta Top
Second reviewer: Dr. Michael Seidl

Date: 11-08-2022

Abstract

Coronavirus disease 2019 or COVID-19 is originated in Hubei province of China in late 2019. Since then, it spread worldwide and caused a worldwide pandemic through most of 2020 and 2021. Gut microbiota has a key role in human health through its protective, trophic, and metabolic actions. Alteration of gut microbiota or gut dysbiosis is found in other virus infection (e.g., hepatitis B, HIV, and influenza). This raises a possibility that COVID-19 might also influence the gut microbiota. Metagenomic analysis is a powerful method to analyze microbiome composition in an environment. In this review, five studies that used metagenomic sequencing to analyze gut microbiome composition in COVID-19 patients were compared. The methods compared include taxonomic profiling, functional annotation, and differential abundance analysis. Based on the approach and tools used, one study differs substantially from the other four studies. This study used protein to protein BLAST for taxonomic profiling, manual alignment to various databases for functional annotation, and non-parametric test of Kruskal-Wallis H and Wilcoxon rank-sum test for differential abundance. In contrast, all four other studies used the biobakery pipeline (MetaPhlAn for taxonomic profiling and HUMAnN for functional annotation) and MaAsLin for the differential abundance analysis of the microbial taxa and pathway. The dissimilarity of the methods between these studies is reflected in the results. The results from each of the four other studies share more agreements with each other even though they are quite different as well. Besides the methods and tools used, the results of these studies are affected by their experimental design (e.g., sample size, sample collections, patients' comorbidities), and the gut microbiota itself is influenced by a lot of other factors (e.g., diet, lifestyle, and medication). It is very difficult to point out whether the differences in the results are caused by the different tools used.

Layman's Summary

Coronavirus disease 2019 or COVID-19 is originated in Hubei province of China in late 2019. Since then, it spread worldwide and caused a worldwide pandemic through most of 2020 and 2021. Microbiota refers to microorganism community of a particular site or habitat. Gut microbiota has a key role in human health because they produce substances important for our body and protect us from infection. Alteration of gut microbiota composition is found in other virus infection (e.g., hepatitis B, HIV, and influenza). This raises a possibility that COVID-19 might also influence the gut microbiota. Metagenomic approach analyzes the whole microbial DNA found in a sample and is a powerful method to analyze microbiome composition in an environment. In this review, five studies that used metagenomic sequencing to analyze gut microbiome composition in COVID-19 patients were compared. The methods compared include: 1) taxonomic profiling, which aims to get the relative abundance of microbial taxa, 2) functional annotation, which give insight into the abundance of biological functions in the microbial community, and 3) differential abundance analysis, which determine the difference in the abundance of microbial taxa and pathways between communities. Based on the approach and tools used, one study differs substantially from the other four studies. This study used protein to protein BLAST for taxonomic profiling, manual alignment to various databases for functional annotation, and Kruskal-Wallis H and Wilcoxon rank-sum test for differential abundance. In contrast, all four other studies used the biobakery pipeline (MetaPhlAn for taxonomic profiling and HUMAnN for functional annotation) and MaAsLin for the differential abundance analysis of the microbial taxa and pathway. The dissimilarity of the methods between these studies is reflected in the results. The results from each of the four other studies share more agreements with each other even though they are quite different as well. Besides the methods and tools used, the results of these studies are affected by their experimental design (e.g., the number of samples, the time of sample collections, patients' additional health condition), and the gut microbiota itself is influenced by a lot of other factors (e.g., diet, lifestyle, and medication). It is very difficult to point out whether the differences in the results are caused by the different tools used.

Introduction

Coronavirus disease 2019 or COVID-19 originated in Hubei province of China in late 2019. Since then, it spread worldwide and caused a worldwide pandemic through most of 2020 and 2021. COVID-19 is caused by a novel betacoronavirus, SARS-CoV-2 (Lake, 2020). This virus is closely related to SARS-CoV (or SARS-CoV-1) that is responsible for the 2002-2004 severe acute respiratory syndrome (SARS) outbreak (Zhou et al., 2020a). It is an enveloped, positive-sense, single-stranded RNA virus that infects lungs epithelial cells. This virus utilizes the angiotensin-converting enzyme 2 (ACE2) receptor to enter the epithelial cell (Zhou et al., 2020b). In addition to lung, kidney and gastrointestinal tract epithelial cells are also expressing the ACE2 receptor and are known to contain SARS-CoV (Harmer et al., 2002; Leung et al., 2003).

SARS-CoV-2 has high rates of transmission, mild to moderate clinical symptoms, with elderly and person with comorbidities (e.g., diabetes melitus, heart disease, and asthma) having a higher risk of severe manifestation (Contini et al., 2020). Mild to moderate symptoms of COVID-19 includes fever, cough, tiredness, anosmia (loss of taste or smell), sore throat, and diarrhea. COVID-19 patients with serious symptoms have difficulties in breathing or shortness of breath and chest pain. Treatments for severely ill patients include support of respiration such as ventilation and corticosteroid treatment using dexamethasone (Wiersinga et al., 2020). With various efforts of social restriction and vaccine development and application, as well as the development of herd immunity, currently the spread of COVID-19 has subsided in a lot of countries.

Role of Microbiota and Gut-lung Axis

The term microbiota refers to microorganism community of a particular site or habitat. Animal body, including human, is one of the habitats of microorganisms. Human gut microbiota consists of 10^{14} microorganisms. This includes bacteria, archaea, viruses, and fungi (Gill et al., 2006). Gut bacteria of healthy individuals is dominated by four phyla, namely Actinobacteria, Firmicutes, Proteobacteria, and Bacteroidetes (Villanueva-Millán et al., 2015).

Gut microbiota has a key role in human health through its protective, trophic, and metabolic actions (Dhar et al., 2020). Microbiota can help the host physiological functions by helping dietary digestion, synthesizing metabolomes (e.g., vitamin K produces by *E. coli* in our gut), and providing protective immunity against pathogens (Wang et al., 2021).

Patients with viral infections, such as hepatitis B virus (Ren et al., 2017), human immunodeficiency virus (Vázquez-Castellanos et al., 2018), and influenza virus (Deriu et al., 2016) frequently observed to have disturbance in their gut microbiome composition. The disruption or alteration of microbiome composition is called dysbiosis. Virus may interact with microbiota and capable of disrupting the microbiome composition, leading to increased inflammation (Ma et al., 2019). Conversely, the altered gut microbiome composition may increase the susceptibility to virus infections, causing more severe clinical symptoms (Hussain et al., 2018).

Gut microbiota is known to affect pulmonary health through a cross-talk between the gut microbiota and the lungs. This process is known to be bidirectional and referred to as the "gut-lung axis" (Keely et al., 2012; Dumas et al., 2018). This, complemented by gut dysbiosis found in other virus infection (Yildiz et al., 2018), raises a possibility that SARS-CoV-2 infection might also influence the gut microbiota.

Metagenomics Approach in Microbiota Analysis

The most widely used method to analyze microbiome composition in an environment is amplicon sequencing. In amplicon sequencing, a common taxonomically informative genomic marker is targeted and amplified by polymerase chain reaction (PCR). In the case of bacteria and archaea, the marker is usually the 16S ribosomal RNA-encoding gene. The marker amplicons are sequenced and analyzed to determine which microbes are present and what their relative abundance is (Sharpton, 2014). This method is powerful but has several limitations. First, the amplification step by PCR gives biases to the presence and relative abundance of the microbial community. This is because the differences in primer affinity across genome, intrinsic features of the genomes (e.g., GC content), and the stochastic nature of PCR experiment could lead to over- or under-amplification of PCR products (Polz and Cavanaugh, 1998). Second, this method is not suitable to study novel or highly diverged microbes since amplicon sequencing depends on taxonomically informative genetic markers, which may not known and hence cannot be amplified for some rare taxa. Third, the widely used 16S rRNA marker gene is a multicopy gene with a strain-specific number of copies, so the accuracy of the microbes' relative abundance will vary (Segata et al., 2012). In addition, the 16S locus can be transferred between distantly related taxa, further decreasing the accuracy of the microbes' relative abundance estimations (Acinas et al., 2004).

Metagenomic sequencing is an alternative microbiome analysis approach that can avoid the limitations of amplicon sequencing. With this tool, the entire nucleotide content from all the organisms within a sample is isolated and sequenced, instead of only targeting a specific gene or locus for amplification. The total DNA is cleaved into tiny fragments and independently sequenced. The results are short DNA sequences, or reads, that are representatives of various genomes present in the sample.

The major advantage of metagenomic sequencing over amplicon sequencing is that it is not only providing insight into the taxonomic composition of the community, but also the biological functions. Metagenomic sequencing analyze reads from the whole genome, meaning that we do not only get taxonomically informative gene like 16S rRNA, but also functionally informative genes that provide insight into biological functions possessed by the microbes.

One of the limitations of metagenomic sequencing is that the cost to perform this method is relatively expensive compared to amplicon sequencing since it sequences the whole DNA in a sample. However, the sequencing cost is decreasing rapidly. According to NHGRI data (Wetterstrand, 2022), the sequencing cost per mega-base (Mb) dropped from $5292.3 in 2001 to $0.006 in 2021. This makes metagenomics more affordable and makes this limitation becomes more irrelevant in the future.

Another limitation of metagenomics is that it has relatively large and complex data which are complicating its analysis. It can be difficult to determine from which genome a read was derived. The computational power that is required for metagenomic analysis is also relatively larger than amplicon analysis. Fortunately, metagenomic software development is advancing rapidly. Metagenomic analysis is becoming easier, faster, and more efficient with the development of new tools (Wang et al., 2021).

However, the advancement of various metagenomic tools also generates confusion for researcher as to what tools are preferable to use. Each tool has its own assumptions and the answer to which tool is the best is highly dependent on the research question. In this review, different tools of metagenomic sequencing used to assess gut microbiota of COVID-19 patients are compared. This involves the taxonomic classification, functional annotation, and differential abundance analysis tools. We compared five studies that used metagenomic sequencing to analyze gut microbiome composition in COVID-19 patients (Zuo et al., 2020; Liu et al., 2021; Yeoh et al., 2021; Liu et al., 2022; Zhang et al., 2022).

Overview of the Reviewed Studies

Zuo et al. (2020) examined three study groups: 15 COVID-19 patients, 6 community-acquired pneumonia (CAP) patients, and 15 healthy individuals as control. Liu et al. (2021) examined COVID-19 symptomatic, asymptomatic, and healthy controls with 10 individuals in each group. The other three studies (Yeoh et al., 2021; Liu et al., 2022; Zhang et al., 2022) only examined two groups, COVID-19 and non-COVID-19, but with much larger sample sizes. All studies except Liu et al. (2021) classified the COVID-19 patient severity into 4 categories: mild, moderate, severe, and critical based on Wu et al. (2020).

Gastrointestinal symptom is observed in all of Liu et al. (2021) patients, while only one patient presents gastrointestinal manifestation in Zuo et al. (2020) study. In Yeoh et al. (2021) and Zhang et al. (2022) study, the proportion of COVID-19 patients that is having gastrointestinal symptoms is 17% and 12% respectively. There is no data available regarding gastrointestinal symptoms of patients in Liu et al. (2022). Almost half of Zuo et al. (2020) COVID-19 patients received antibiotic treatment, while 34% and 23.6% patients received antibiotics in Yeoh et al. (2021) and Liu et al. (2022) studies respectively. In Zhang et al. (2022) study, all COVID-19 patients do not receive any antibiotic treatments.

Stool samples are collected during hospitalization in all studies, with Liu et al. (2022) also collected the samples at 1 month and 6 months after discharge, Zhang et al. (2022) collected the sample beyond 1 month after discharge, and Yeoh et al. (2021) collected the sample up to 30 days after clearance of SARS-CoV-2 based on RT-qPCR result of nasopharyngeal swab. Details regarding the subjects and sample collections of the five reviewed studies are shown in Table 1.

Zuo et al. (2020) and Yeoh et al. (2021) measured SARS-CoV-2 viral load in the stool samples to study its correlation with gut microbiota. In addition, Yeoh et al. (2021) also measured cytokines and chemokines concentration from blood samples. Meanwhile, Zhang et al. (2022) performed fecal metabolites measurements which include short chain fatty acids (SCFAs) and

L-isoleucine measurements. SCFAs measured include acetic, propionic, isobutyric, butyric, isovaleric, valeric, and hexanoic acid.

Table 1. Subject and sample collection details of the reviewed studies

	Zuo et al. (2020)	Liu et al. (2021)	Yeoh et al. (2021)	Liu et al. (2022)	Zhang et al. (2022)
Subject	• 15 COVID-19 patients; • 6 CAP patients; • 10 healthy individuals	• 10 COVID-19 symptomatic patients; • 10 COVID-19 asymptomatic patients; • 78 healthy individuals	• 100 COVID-19 patients; • 78 healthy individuals	• 106 COVID-19 patients; • 68 healthy individuals	• 66 COVID-19 patients; • 70 healthy individuals
Female	• 53% in COVID-19; • 33% in CAP; • 40% in healthy control	• 40% in symptomatic; • 40% in asymptomatic; • 30% in healthy control	• 47% in COVID-19; • 58% in healthy control	53% in COVID-19	• 41% in COVID-19; • 59% in healthy control
Age	Median (IQR): • 55 (44-67.5) in COVID-19; • 50 (44-65) in CAP; • 48 (45-48) in healthy control	Mean (± SD): • 39 (± 11) in symptomatic; • 36 (± 11) in asymptomatic; • 44 (± 14) in healthy control	Mean (± SD): • 36.4 (± 18.7) in COVID-19; • 45.5 (± 13.3) in healthy control	Median (IQR): 48.3 (33-62) in COVID-19	Mean (± SD): • 42.6 (± 19.0) in COVID-19 • 45.8 (± 13.7) in healthy control
Severity of COVID-19 patients	• Mild: 1 (7%) • Moderate: 9 (60%) • Severe: 3 (20%) • Critical: 2 (13%)	NA	• Mild: 47 (47%) • Moderate: 45 (45%) • Severe: 5 (5%) • Critical: 3 (3%)	• Asymptomatic: 4 (3.8%) • Mild: 31 (29.2%) • Moderate: 55 (51.9%) • Severe: 10 (9.4%) • Critical: 6 (5.7%)	• Mild: 31 (47%) • Moderate: 16 (24.2%) • Severe: 15 (22.7%) • Critical: 4 (6.1%)
Comorbidities	40% in COVID-19; 100% in pneumonia	No comorbidities	31% in COVID-19: • hypertension (11%) • hyperlipidaemia (4%) • diabetes (2%) • heart condition (2%) • allergic disorder (7%) • HIV (3%) • asthma (2%) 28% in healthy control: • (hypertension (11%) • allergic disorders (15%) • asthma (2%)	43% in COVID-19: • hypertension (17%) • diabetes (15%) • hyperlipidaemia (11%)	27% in COVID-19: • hypertension (9%) • hyperlipidaemia (11%) • asthma (4.5%) • HBsAg (4,5%) • diabetes (4.5%) • allergic rhinitis (3%) • HIV (3%) • heart disease (1.5%) • eczema (1.5%) 29% in healthy control: • hypertension (14%) • eczema (3%) • allergic rhinitis (10%) • asthma (3%) • bowel disease (7%) • hemorrhoids (6%)
Antibiotics use of COVID-19 patients	47%	NA	34%	23.6%	0%
Sample collection	2-3 times per week during hospitalization	Once	Once (all); Serially up to 30 days after tested negative for SARS-CoV-2 (27 patients)	3 times (at admission, 1 month, and 6 months after discharge	Up to 6 times during hospitalization and beyond 1 month after discharge

Taxonomic Classification

Strategies in Taxonomic Classification

The most common goal in metagenomic analysis is to know what microbial species or taxa present in a community. This information is not only useful in and of itself but also for comparative studies where similarity between two or more communities is assessed. Taxonomic classification or taxonomic profiling aims to get relative abundances of taxa within metagenomics data (Ye et al., 2019). Taxonomic classification may also provide insight into the biological function and condition of the community if it contains members of functionally described taxa. For example, the presence of opportunistic taxa suggests that the community is in the state of disruption. Taxonomic classification is usually performed by three non-

exclusive methods: 1) marker gene analysis, 2) binning, and 3) assembly (Sharpton, 2014), which will be discussed in more detail below.

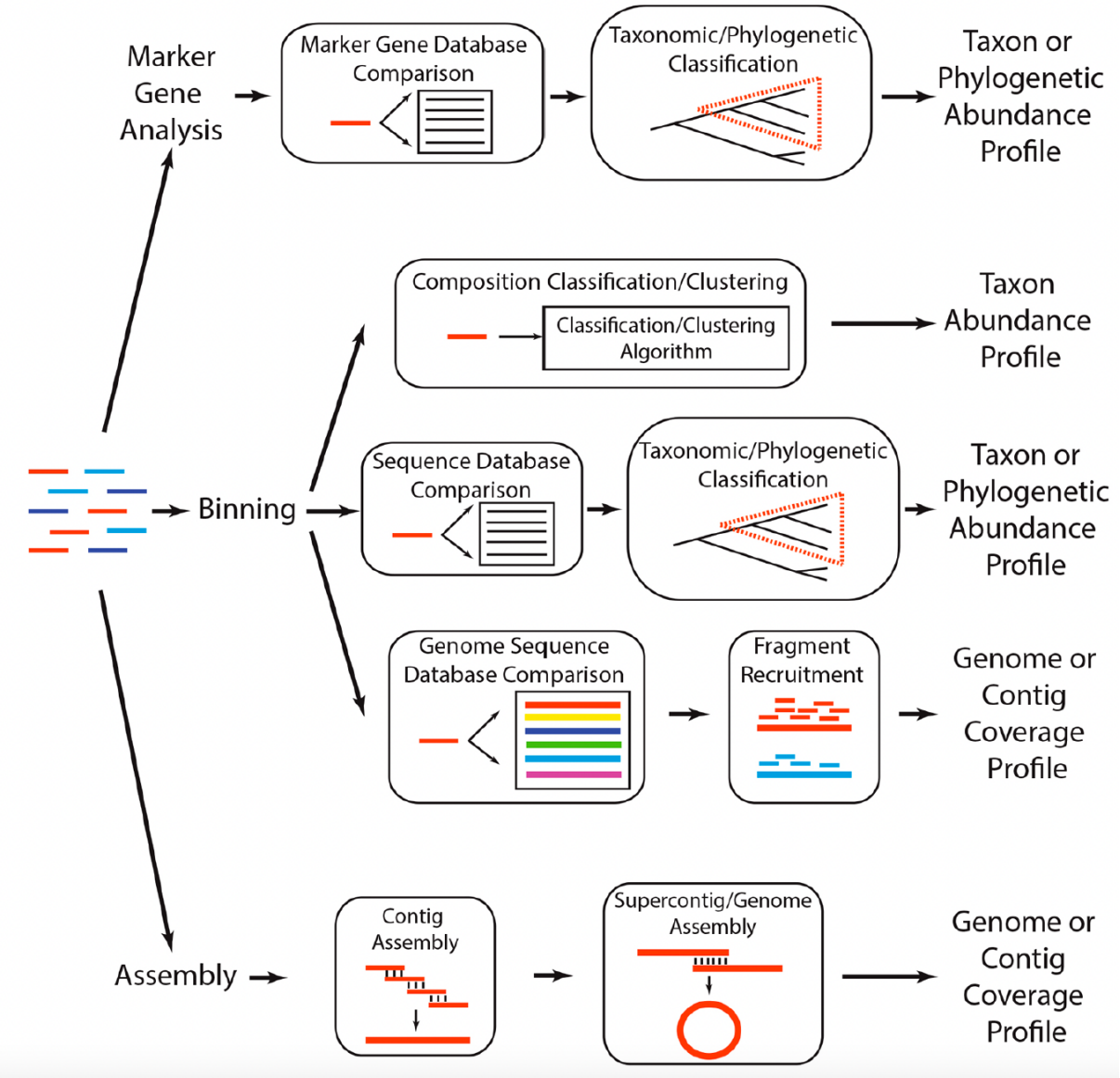


Figure 1. Strategies of taxonomic classification in metagenomic analysis (adapted from Sharpton, 2014)

Marker gene analysis compares metagenomic reads to a database of taxonomically informative gene markers. These gene markers are used as classifiers because they are strongly conserved within a taxon's genomes and are unique enough to not having substantial similarity with sequences outside the taxon (Segata et al., 2012). Marker gene analysis is computationally efficient because it does not align every read to every available genome. MetaPhlAn (https://github.com/biobakery/MetaPhlAn) is one of the most used taxonomic classifiers that utilizes marker gene analysis approach, which is developed by researchers from Harvard School of Public Health and University of Trento (Beghini et al., 2021). They pre-determined clade-specific marker genes that can assign reads to microbial taxa unambiguously, accurately, and efficiently (Beghini et al., 2021). MetaPhyler is another tool that uses marker gene approach (Liu et al., 2011)

Binning aims to assign individual sequence reads to taxonomic groups (Meyer et al., 2022). By clustering the reads, binning reduces the complexity of the data. It also provides insight into the numbers and types of taxa in the community. In addition, it may also provide insight into the presence of novel taxa, depending on the method used. Binning commonly performed in three different approaches: 1) sequence compositional binning, which utilizes sequence characteristics such as GC content and tetramer frequency, 2) sequence similarity binning, which uses sequence similarity of metagenomic sequences to a database, and 3) fragment recruitment binning, which maps reads to genome sequences that exhibit nearly identical alignments (Sharpton et al., 2014). Tools capable of doing binning process includes PhyloPythia (McHardy et al., 2007), CONCOCT (Alneberg et al., 2014), and MetaBAT (Kang et al., 2019) that used compositional approach, MEGAN (Huson et al., 2011) and MaxBin (Wu et al., 2016) which used sequence similarity approach, and MOSAIK that used fragment recruitment approach (Lee et al., 2013).

Assembly merges metagenomic reads from the same genome into a single contiguous sequence called contig. Assembly can be done with two strategies: reference-guided and *de novo* assembly. In reference-guided assembly, reads are aligned to the reference genomes to build contigs. This method performance depends on the availability of the reference genome and the stringency of the reads alignment. *De novo* assembly does not need a reference and mainly build upon the traditional de Bruijn graph approach. This approach basically breaks up all reads into shorter sequences of length k (k-mers), and a graph is created by sequential k-mers in the reads. In this approach, the k-mers act as nodes while the reads act as edges (Compeau et al., 2011). Contig simplify bioinformatic analysis compared to unassembled metagenomic reads. The biggest challenge of assembly is to minimize the risk of chimera sequence generation. To mitigate this, researchers often bin reads prior to assembly (Sharpton et al., 2014). Tools that are commonly used for assembly include HipMer (Hofmeyr et al., 2020), SPAdes (Nurk et al., 2017), GATB (Drezen et al., 2014) and Megahit (Li et al., 2015).

Comparison of Taxonomic Classifier Used in the Reviewed Studies

In 4/5 studies, the marker gene analysis approach was used without a binning and assembly process (Zuo et al., 2020; Yeoh et al., 2021; Liu et al., 2022; Zhang et al., 2022) (Table 2). In these studies, MetaPhlAn2 was used, except for Liu et al. (2022) that used MetaPhlAn3.

Table 2. Taxonomic classification approach used by the reviewed studies

	Zuo et al. (2020)	**Liu et al. (2021)**	**Yeoh et al. (2021)**	**Liu et al. (2022)**	**Zhang et al. (2022)**
Assembly	-	Megahit	-	-	-
Gene prediction	-	MetaGene	-	-	-
Taxonomic classification	MetaPhlAn2 (V2.9)	BLASTp towards NCBI non-redundant database	MetaPhlAn2 (V2.7.7)	MetaPhlAn3 V3.0.5	MetaPhlAn2

MetaPhlAn mapped each metagenomic reads to a pre-defined clade-specific marker catalog using bowtie2 (Langmead and Salzberg, 2012). The original catalog contains a total of 400,141 genes for all taxonomic unit, spanning 1,221 species with 231 markers per species (s.d. 107) and more than 115,000 markers at higher taxonomic levels. MetaPhlAn2 expanded the catalog to ~1 million markers (184 ± 45 for each bacterial species) from a total of more than 7,500 species. MetaPhlAn2 also incorporates subspecies markers which enable strain-level analyses (Truong et al., 2015). MetaPhlAn3 further expand the catalog, consisting of 1.1 million marker genes (84 ± 47 markers per species) spanning 13.5 thousand species (Beghini et al., 2021).

These markers were selected from pan-proteome database of UniRef90. In short, UniRef90 is a UniprotKB database that is clustered at 90% sequence identity to reduce database redundancy (Suzek et al., 2015). All proteins with length between 150 and 1500 amino acids are used for marker discovery. The markers are selected based on two values: "coreness" value and "uniqueness" value. The "coreness" value represents how conserved is the marker within the clades, and the "uniqueness" value represents how often the marker sequence is shared with other clades. A score function defined by "coreness" value and "uniqueness" value is created to determine the quality of the markers (Beghini et al., 2021).

On the other hand, Liu et al. (2021) used a different approach in classifying their metagenomic data. First, they performed reads assembly using Megahit (Li et al., 2015). Then, *de novo* gene prediction is performed on the assembled contigs using MetaGene (Noguchi et al., 2006). The genes are then clustered using CD-HIT and aligned using SOAPaligner to select gene representatives and calculate the gene's relative abundances. The taxonomic classification is done in protein level by performing BLASTp (Version 2.2.28+) on the translated gene sequences to NCBI non-redundant (NR) sequence database.

Megahit is a *de novo* assembler that makes use of succinct de Bruijn graphs (SdBG). SdBG is a compressed representation of de Bruijn graphs. This tool implemented multiple k-mer size strategy, where it iteratively builds multiple SdBGs from a small k to a large k. A small k-mer size is useful in filtering erroneous edges and filling gaps in low coverage regions, while a large k-mer size can resolve repeats (Li et al., 2015).

BLAST is considered as the "gold standard" for sequence comparison. It offers high sensitivity (the proportion of the total number of sequences assigned correctly) and precision (the proportion of assigned sequences assigned correctly), but in the price of computational resource requirements. Protein to protein alignment like BLASTp also provides more sensitivity towards novel and highly variable sequence compared to DNA to DNA alignment because the lower rate of mutation of amino acid compared to nucleotide (Ye et al., 2019).

Different Taxonomic Classification Approach is Reflected in the Results

All studies produce fairly different result of differentially abundance taxa between COVID-19 patients and healthy individuals with some similarity. Bacteria of genus *Bacteroides* is found to be enriched in COVID-19 patients in all studies except Liu et al. (2021). However, the *Bacteroides* bacteria enriched in COVID-19 patients differ in species level between studies. *B. nordii* is differentially abundant in Zuo et al. (2020), *B. dorei* and *B. caccae* is significantly

enriched when antibiotic intake was not considered and was considered as a co-variant respectively in Yeoh et al. (2021), *Bacteroides* in genus level is differentially abundant in Liu et al. (2022), and *B. ovatus*, *B. dorei*, and *B. thetaiotaomicron* is differentially abundant in Zhang et al. (2022) (Table 3). In addition, *B. vulgatus* is associated with Post-Acute COVID Syndrome (PACS) in Yeoh et al. (2021). Interestingly, *B. dorei*, *B. thetaiotaomicron*, *B. massiliensis*, and *B. ovatus* are negatively associated with SARS-CoV-2 viral load in Zuo et al. (2020) study, contradicting Zhang et al. (2022) study. Meanwhile, Liu et al. (2021) results barely share any similarity with other studies. The only similarity is the higher level of *Erysipelotrichaceae* and *Actinomyces* found in asymptomatic cases based on LEfSE analysis, which are also found in COVID-19 patients in Zuo et al. (2020).

Table 3. Significantly different taxa in the reviewed studies (bold taxa indicates multiple findings)

	Zuo et al. (2020)	Liu et al. (2021)	Yeoh et al. (2021)	Liu et al. (2022)	Zhang et al. (2022)
Enriched in COVID-19 patients compared to healthy control	• *Clostridium hathewayi* • *Actinomyces viscosus* • ***Bacteroides*** *nordii*	Phylum level: • Candidatus_ Saccharibacteria • Synergistetes • Candidatus_ Gottesmanbacteria • Candidatus_ Moranbacteria Genus level: • *Lachnoclostridium* • *unclassified _f_Erysipelotrichaceae* • *Tyzzerella*	Phylum level: **Bacteroidetes** Species level: • *Ruminococcus gnavus* • *R. torques* • ***Bacteroidetes*** *dorei* Antibiotics examined as co-variant: • Parabacteroides • *Sutterella wadsworthensis* • ***Bacteroides*** *caccae*	• ***Bacteroides*** • *Blautia*	• ***Bacteroides*** *ovatus* • ***Bacteroides*** *dorei* • ***Bacteroides*** *thetaiotaomicron*
Depleted in COVID-19 patients compared to healthy control	• ***Eubacterium*** *ventriosum* • ***Faecalibacterium prausnitzii*** • ***Roseburia*** • ***Lachnospiraceae***	Phylum level: Fibrobacteres	Phylum level: Actinobacteria Species level: • ***Bifidobacterium adolescentis*** • ***Faecalibacterium prausnitzii*** • ***Eubacterium*** *rectale* Antibiotics examined as co-variant: • *Adlercreutzia equolifaciens* • *Dorea formicigenerans* • *Clostridium leptum*	• ***Ruminococcus*** • ***Bifidobacterium*** • *Collinsella* • **Lachnospiraceae** • ***Roseburia*** • Fusicatenibacter • ***Faecalibacterium***	• ***Bifidobacterium adolescentis*** • ***Ruminococcus*** *bromii* • ***F. prausnitzii***
Positively correlated with disease severity	• *Clostridium ramosum* • *C. hathewayi* • *Coprobacillus*			No association found between gut microbiota composition and disease severity	
Negatively correlated with disease severity	• *Alistipes onderdonkii* • *B. ovatus* • ***F. prausnitzii***		• **F. *praunitzii*** • ***Bifidobacterium*** *bifidum*		• ***Bifidobacterium*** *adolescentis* • ***F. prausnitzii***

One bacterial species, *Faecalibacterium prausnitzii* is found to be depleted in COVID-19 patients compared to healthy control in all studies except Liu et al. (2021) (Table 3). This bacterium is also negatively correlated to disease severity (Zuo et al., 2020; Yeoh et al., 2021; Zhang et al., 2022) and PACS (Liu et al., 2022). In addition, *Bifidobacterium adolescentis* is found in two studies to be depleted in COVID-19 patients (Yeoh et al., 2021; Zhang et al., 2022), while in genus level, Liu et al. (2022) also found *Bifidobacterium* to be depleted in COVID-19 patients (Table 3). Liu et al. (2021) does not have any intersecting taxa with any

studies for COVID-19 depleted microbes, where they found Fibrobacteres in phylum level to be depleted in COVID-19 patients.

Functional Annotation

Metagenomic Functional Annotation Workflow

As mentioned above, metagenomic analysis could not only provides the information of taxonomic profile of a microbial community, but also provides information about their biological functions. Generally, metagenome functional annotation comprises of two steps: gene prediction and gene annotation (Figure 2).

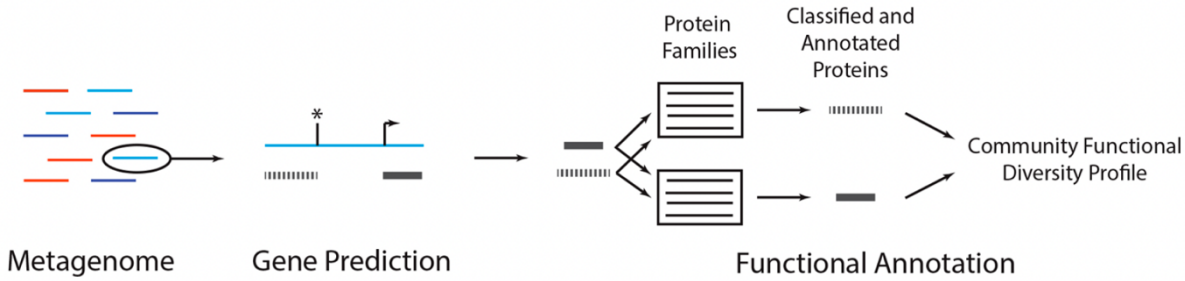


Figure 2. A metagenomic functional annotation workflow (adapted from Sharpton, 2014)

Gene prediction determines the presence of coding sequences in metagenomic reads. This process can be done on assembled or unassembled metagenomic sequences. Gene prediction can be performed by mapping the translated metagenomic sequence to a protein database. Because the coding regions are not known, one needs to translate all six possible protein coding frames and compares each resulting amino acid sequence to the protein database. Sequence translation tools such as Transeq (Madeira et al., 2022) could be used for this process. The protein sequence alignment generally performed using BLASTp. This method expectedly requires high computational resource and time considering the high amount of translated sequence and the full alignment to a protein sequence database (Sharpton et al., 2014).

Another way to predict coding sequences is *de novo* gene prediction. This method is not based on homology so it can potentially identify novel genes. In this method, genes are predicted using gene prediction models which evaluates various properties of microbial genes such as gene length, codon usage, and GC bias. This model will find all possible open reading frame (ORF) – a section of sequence started with start codon and ended with stop codon – and determine whether the ORF is coding for protein based on the properties mentioned above (Noguchi et al., 2006). This method may require a fair bit of time and computer resource, but it is generally faster than translating the reads in 6-frame and do protein alignment with the database.

After the genes in metagenomic data are predicted, gene annotation is performed to predict their functionality. It is difficult to actually determine the function of a protein; it is a research on its own. So, the general approach is to assign each predicted gene to protein families by

doing sequence alignment with a database (Sharpton et al., 2014). Protein families are groups of evolutionary related proteins. Related proteins generally have high similarity in their sequences and therefore, are thought to have similar biological functions. The metagenomic predicted protein function thus can be inferred based on its sequence similarity with protein families.

Different databases offer different flavors to the functional annotation. Kyoto Encyclopedia of Genes and Genomes (KEGG) offers comprehensive metabolic pathway modules for protein families (Kanehisa et al., 2014). MetaCyc also has highly curated and well-described metabolic pathways (Caspi et al., 2014). EggNOG provides protein orthologs groups non-supervised database which is claimed to has a higher precision than traditional homology searches like BLAST (Huerta-Cepas et al., 2016). Instead of sequence similarity, Pfam (Mistry et al., 2020) protein family database is based on hidden Markov models (HMMs) of protein domains. The HMMs database offers more sensitivity in identifying more distantly related or diverged members of a family compared to the sequence similarity model used by other databases.

Comparison of Functional Annotation Pathways Used in the Reviewed Studies

From five studies reviewed in this research, only three of them performed functional annotation of the metagenomic data. Zuo et al. (2020) and Yeoh et al. (2021) did not perform functional annotation in their study. The summary of functional annotation approach used by the other three studies is shown in Table 4.

Liu et al. (2021) performed a set of functional annotation steps that includes gene prediction, clustering, and gene annotation. They performed gene prediction on the assembled contigs using MetaGene. MetaGene is a *de novo* gene prediction tool that takes into account various properties of a gene including codon frequencies (and di-codon frequencies), distribution frequency of open reading frame (ORF) lengths, distance from the leftmost start codons, and distances between neighboring ORFs to differentiate between protein-coding ORFs and random ORFs (Noguchi et al., 2006).

Predicted genes are clustered using CD-HIT with the threshold of 95% sequence identity and 90% coverage. The longest sequences in each cluster were selected as the representative sequence to construct non-redundant gene catalog. The reads were then mapped to the representative sequence using SOAPaligner (95% identity threshold) to calculate the abundance.

Table 4. Functional annotation approach used by three studies

	Zuo et al. (2020)	Liu et al. (2021)	Yeoh et al. (2021)	Liu et al. (2022)	Zhang et al. (2022)
Functional annotation	NA	BLAST towards eggNOG, KEGG, Hmmscan, ARDB, VFDB, CARD, and GO databases	NA	HUMAnN3	HUMAnN2

The representative sequences were functionally annotated by BLAST into various databases including eggNOG, KEGG, Hmmscan (for carbohydrate-active enzymes), ARDB (antibiotic

resistance database), VFDB (virulent factor database), CARD (comprehensive antibiotic resistance database), and GO (gene ontology).

In the studies by Liu et al. (2022) and Zhang et al. (2022), the HUMAnN (version 3 and 2, respectively) (Beghini et al., 2021; Franzosa et al., 2018) pipeline was used for the functional annotation and assignment of the metagenomic reads to metabolic pathways. HUMAnN2 performs functional annotation of metagenomic reads with a "tiered search" strategy. Microbial species identification of MetaPhlAn2 is the first-tier search. HUMAnN2 then collects the identified species pangenomes that are functionally annotated to construct a database and uses it as a reference to map all the metagenomic reads. This mapping process is the second-tier search. In the third tier, all reads that are not mapped to the pangenomes are aligned towards a UniprotKB protein database. The mapped reads are then used to calculate the relative abundance of protein families, which can be linked to other databases such as KEGG, eggNOG, and Pfam to get the information of the biological pathways and functional groups. By default, MetaCyc is used in HUMAnN2 to reconstruct and quantify metabolic pathways in the microbial community (Franzosa et al., 2018).

HUMAnN3 is an updated and improved version of HUMAnN2. It uses an updated ChocoPhlAn 3 database to construct the pangenome database in metagenomic functional annotation process. HUMAnN3 implements several fine-tuning to improve its performance, including bowtie2 and DIAMOND search parameters adjustment for the second-tier and third-tier steps respectively and tuning on the human-like synthetic metagenome data to reduce overfitting. It also implements a coverage filtering in assigning reads to pangenomes that improves its specificity (Beghini et al., 2021).

Table 5. Significantly different biological pathway in the reviewed studies (bold pathway indicates multiple findings)

Liu et al. (2021)	Liu et al. (2022)	Zhang et al. (2022)
Enriched in asymptomatic cases: • Pentose phosphate pathway • Insulin signaling • RNA polymerase • Phenylalanine metabolism • Ascorbate and aldarate metabolism • AMPK signaling	Enriched in COVID-19 patients: • **Urea cycle** • L-citrulline biosynthesis • **L-ornithine biosynthesis** • All-trans-farnesol biosynthesis • Pyruvate fermentation to butanoate • L-histidine degradation • Fatty acid biosynthesis • L-isoleucine biosynthesis	Enriched in COVID-19 patients: • **Urea cycle** • Peptidoglycan biosynthesis • **L-ornithine biosynthesis** • Sugar nucleotide biosynthesis • Heme biosynthesis • Purine nucleotide biosynthesis
	Depleted in COVID-19 patients: • Phytate degradation • Sulfur oxidation • L-proline biosynthesis • Clostridium acetobutylicum acidogenic fermentation	Depleted in COVID-19 patients: • Carbohydrate degradation • Bifidobacterium shunt (acetic acid biosynthesis) • Fatty acid degradation • Amino acid biosynthesis (L-isoleucine, L-methionine, L-histidine, branched amino acids) • NAD biosynthesis • Purine nucleotide biosynthesis

Functional Annotation Results Comparison

Both Liu et al. (2022) and Zhang et al. (2022) found urea cycle pathway and L-ornithine biosynthesis pathway to be significantly enriched in COVID-19 patients compared to healthy

control (Table 5). Zhang et al. (2022) found 19 pathways that are significantly depleted in COVID-19 patients compared to healthy control, with 7 of them are related to carbohydrate degradation. Liu et al. (2022) found 6 pathways depleted in COVID-19 patients but they are not found in Zhang et al. (2022) findings. Meanwhile, Liu et al. (2021) found seven pathways that are significantly enriched in asymptomatic individuals compared to healthy control, but none of them overlaps with the pathways found in Liu et al. (2022) and Zhang et al. (2022) (Table 5).

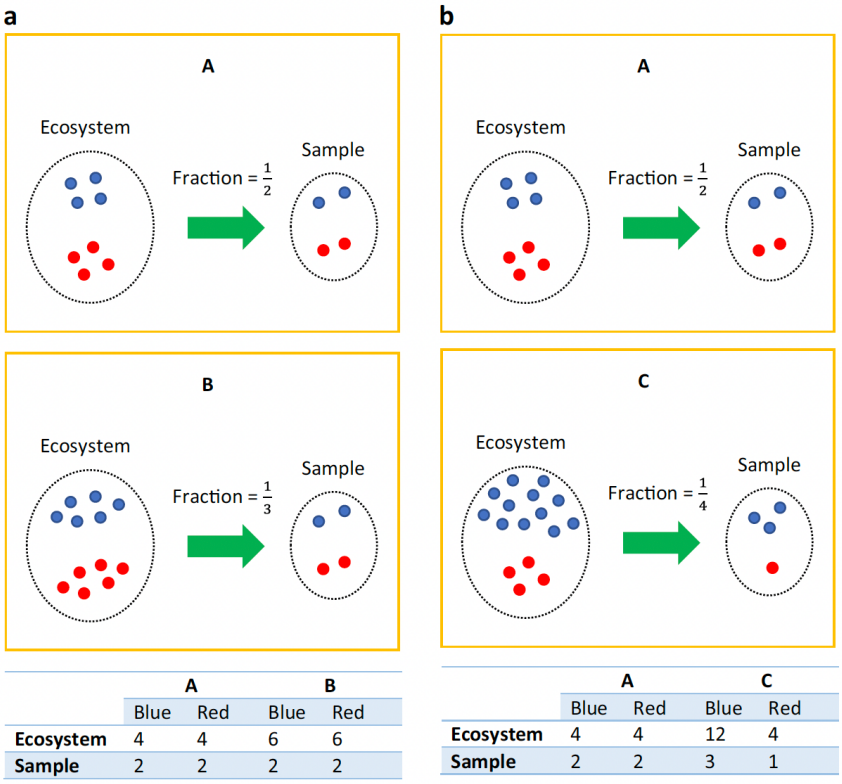
Differential Abundance Analysis

The Importance of Differential Abundance Analysis

Differential abundance analysis aims to determine whether the abundances of a microbial taxa between two ecosystems are different. This could also be applied to determine differentially abundant pathway in metagenomic functional analysis. Using the right approach in differential abundance analysis is then as important as choosing the right taxonomic classification and functional annotation method when comparing two or more microbial communities. Unfortunately, there is little consensus on how to implement differential abundance analysis in microbial data. Dozens of tools exist with different approaches in data input, normalization, and general assumptions of the data distribution. Nearing et al. (2022) compare the performance of 14 differential abundance testing methods and found that they produce very different results.

The most important challenge in differential abundance analysis is normalization. Normalization is needed because each sample has different sampling fractions. Sampling fraction is the ratio of observed abundance to unobserved absolute abundance of each taxon. While the observed abundance is known from the experiment, we do not know the sampling fraction nor the absolute abundance. A taxon could seem to be more abundant in sample A compared to sample B based on our observed abundance data, but it might just because sampling fraction of sample A is much bigger than sample B. The absolute abundance of that taxa in sample A and B might not actually be different, thus producing false positive result (Figure 3).

Rarefying is one of the methods of normalization. This procedure aims to deal with differences in library sizes. First, the minimum library size is determined – samples with library size smaller than the minimum will be discarded. Second, sample with library size larger than the minimum are subsampled without replacement so that all samples have the minimum library size. Minimum library size can be selected based on rarefaction curves, which represent diversity as a function of library size. Library size where the diversity starts stagnating (approach a slope of zero) is considered a good minimum library size because the diversity has been fully observed (Lin and Peddada, 2020). While rarefying is useful in omitting biases from different library sizes, it has its own caveat. The omission of available valid data, the introduction of uncertainty in subsampling step, the arbitrary selection of minimum library size, and the challenges in estimating over-dispersion parameter, are several concerns raised regarding this method (Lin and Peddada, 2020).



	A		B	
	Blue	Red	Blue	Red
Ecosystem	4	4	6	6
Sample	2	2	2	2

	A		C	
	Blue	Red	Blue	Red
Ecosystem	4	4	12	4
Sample	2	2	3	1

Figure 3. Representation of sampling fraction (adapted from Lin and Peddada, 2020).
a) False negative introduced by sampling fraction. Taxa in subject A and B have similar observed abundance but actually have different unobserved abundance in the two ecosystems, because of the difference of sampling fraction. b) False positive introduced by sampling fraction. When comparing subject A and C, both the blue and red taxa seem to be differentially abundant. In fact, only the blue taxon is differentially abundant. The red taxon has the same unobserved abundance in A and C.

Scaling is another popular method of normalization that aims to solve the sampling fraction bias. Basically, scaling divides the observed abundance by a "scaling factor". Comparing with the relationship between observed abundance, sampling factor, and absolute abundance described above, a scaling factor that is closest to the unknown sampling factor is the most ideal. Several commonly used scaling methods include Total-Sum Scaling (TSS) that is used in MaAsLin (Morgan et al., 2012; Morgan et al., 2015), Cumulative-Sum Scaling (CSS) which is used in MetagenomeSeq (Paulson et al., 2013), Median normalization (MED) that is implemented in DESeq2 (Love et al., 2014), Upper Quartile normalization (UQ) in edgeR (Robinson and Oshlack, 2010), and Trimmed Mean of M-values (TMM) in Wrench (Kumar et al., 2018).

Microbial abundance data is a type of compositional data, which is defined as proportions of some whole. This type of data could be subjected to isomorphism transformation of log-ratio, which can eliminate the effect of sampling fraction. In this method, log-ratios of all taxa are obtained with respect to a common reference taxon of all taxa. This way, the bias of sampling fraction is intrinsically eliminated. This log-transformation is called the additive log-ratio (alr) transformation. Other than using a particular taxon, the log-ratio based method can also be performed by using the center of mass of all taxa as the reference. This transformation is called centered log-ratio (clr) transformation (Lin and Peddada, 2020).

Besides normalization, an aspect which is in disagreement between different differential abundance tools is the statistical distribution used in modeling the observed abundance data. DESeq2 (Love et al., 2014) and edgeR (Robinson and Oshlack, 2010) use negative binomial (NB) distribution which is inspired by transcriptomics data. MetagenomeSeq (Paulson et al., 2013) is based on zero-inflated Gaussian (ZIG) model. In this model, zeros are classified to sampling zeros and structural zeros, where sampling zeros is the effect of insufficient library size and structural zeros is caused by the nature of the ecosystem itself. ALDEx2 (Fernandes et al., 2014) use Dirichlet-multinomial distribution, while Microbiome Multivariable Associations with Linear Models (MaAsLin; Morgan et al., 2012; Morgan et al., 2015) and limma voom (Law et al., 2014; Ritchie et al., 2015) are based on normal distribution.

Comparison of Differential Abundance Analysis Tools Used in the Reviewed Studies

All our reviewed studies except Liu et al. (2021) used MaAsLin to determine the differential abundance (Table 6). MaAsLin is based on applications of arcsine square root-transformed (AST) linear model and use TSS as normalization (Morgan et al., 2012; Morgan et al., 2015). The AST model is commonly used and has become a standard in analyzing proportional data in ecology for long (Warton and Hui, 2011). This model is thought to resemble the microbial abundance data. However, TSS scaling used in this tool may be too simplistic and may introduce bias (Weiss et al., 2017). TSS scaled each taxon by the sample's library size. In other words, it transforms the observed abundance into relative abundance. This scaling does not resolve the sampling factor issue and can introduce bias in differential abundance estimates because change in the abundance of one taxon can influence the relative abundances of all taxa.

On the other hand, Liu et al. (2021) used non-parametric test of Kruskal-Wallis H (for multiple sample classes) and Wilcoxon rank-sum test (for two sample classes) to analyze the differential abundance. Contrary to the parametric test, non-parametric test is not based on assumptions about the data distribution. This is actually not preferred for a compositional-type data like microbial abundance.

Table 6. Differential abundance analysis tools used by the reviewed studies

	Zuo et al. (2020)	Liu et al. (2021)	Yeoh et al. (2021)	Liu et al. (2022)	Zhang et al. (2022)
Differential abundance tool	MaAsLin	Kruskal-Wallis H test or Wilcoxon rank-sum test; LEfSe	MaAsLin; LEfSe	MaAsLin; LEfSe	MaAsLin

In addition, Yeoh et al. (2021), Liu et al. (2022), and Liu et al. (2021) also performed Linear Discriminant Analysis Effect Size (LEfSe). Rather than determining the difference in abundance of microbial taxa between samples, LEfSe is more focused on analyzing the relationship between microbial taxa and a phenotype. LEfSe aims to quantify the magnitude of the effect size of such association.

LEfSe uses rarefied data as an input, so it eliminates biases from different library sizes but introduces uncertainty in the subsampling. The scaling used in LEfSe is TSS scaling, which as mentioned above, is too simplistic and may introduce bias. LEfSe performs Linear Discriminant Analysis (LDA) to calculate the effect size of each taxon to a phenotype (e.g.,

severity). The observed abundance of taxa acts as the independent variable and the phenotype features as the dependent variable. LDA assumes the independent variable to be normally distributed (McLachlan, 2004).

Discussion

Since the emergence of COVID-19 in 2019, plenteous amount of research were conducted in various aspects of the virus and disease, including the putative relation between the virus infection and gut microbiome. These research differ in their overall design as well as the approaches and tools used to analyze the microbiome. This review aims to compare metagenomics analysis tools that were used to analyze gut microbiome of COVID-19 patients.

From the choice of taxonomic classification, functional annotation, and differential abundance approach performed in these studies, the methods used by Liu et al. (2021) was very different from the other four studies. In taxonomic classification, Liu et al. (2021) used protein to protein BLAST, while the other studies used marker gene analysis using MetaPhlAn. Liu et al. (2021) used their own pipeline in annotating biological function while other studies used HUMAnN. Finally, Liu et al. (2021) used non-parametric test of Kruskal-Wallis H and Wilcoxon rank-sum test while all other studies used MaAsLin.

MetaPhlAn is a very effective and efficient tool for taxonomic profiling because of their marker gene approach. With its growing database (~13.5 thousand species in version 3), MetaPhlAn provides fast classification with high accuracy. MetaPhlAn2, along with mOTUs 2.5.1 (Milanese et al., 2019), performed best in all communities tested in Meyer et al. (2022). MetaPhlAn3 improved in accuracy compared to MetaPhlAn2 in analyzing human and murine gastrointestinal metagenomes (Beghini et al., 2021). On the other hand, the choice of Liu et al. (2021) to do the taxonomic classification on protein level is disconcerting because it may introduce unnecessary bottlenecks in the process of assembly and gene prediction. The advantage of protein to protein alignment is it provides more sensitivity towards more distantly related and novel proteins, but that advantage is not very useful in comparing a well-studied gut microbiome composition.

All studies except Liu et al. (2022) found *Bacteroides* species to be enriched in COVID-19 patients. *Bacteroides* species, including *B. dorei*, *B. thetaiotaomicron*, *B. massiliensis*, and *B. ovatus* are known to downregulate the ACE2 expression in murine colon (Kalantar-Zadeh, 2020), which will alleviate SARS-CoV-2 replication. In addition, *B. vulgatus* and *B. dorei* is known to suppress pro-inflammatory immune response and can be used as probiotic treatment in influenza-infected and atherosclerosis mice (Song et al., 2022; Yoshida et al., 2018). These research support Zuo et al. (2020) finding that *B. dorei*, *B. thetaiotaomicron*, *B. massiliensis*, and *B. ovatus* is negatively correlated with disease severity. However, *B. dorei* and *B. vulgatus* are also found to be enriched in several inflammatory gut diseases such as irritable bowel disease and ulcerative colitis (Davis-Richardson et al., 2014), which supports Zhang et al. (2022) finding about the enrichment of those bacteria in COVID-19 patients.

The consistent finding of *F. prausnitzii* as the depleted bacteria in COVID-19 patients is very important. *F. prausnitzii* is a commensal bacterium in human gut that is known to have immunomodulatory effect and contribute to host defense. *F. prausnitzii* helps downregulate

inflammatory response by inhibiting NF-kB pathway and interfering the synthesis and suppressing the secretion of interleukin-8, a pro-inflammatory chemokine (Breyner et al., 2017; Ferreira-Halder et. al, 2017). In addition, *F. prausnitzii* produces short chain fatty acid molecules (SCFAs) such as butyrate, propionate, and acetate which are associated with the capacity to induce interleukin-10, an anti-inflammatory cytokine (Xu et al., 2020), alter chemotaxis and phagocytosis, and have anti-microbial and anti-inflammatory effects (Yao 2020). Furthermore, butyrate is known to prevent translocation and circulation of gut endotoxin and bacteria, thus reducing systemic inflammatory response (Geirnaert et al., 2017). This theory of *F. prausnitzii*'s role is supported by Zhang et al. (2022) study which found that SCFA depletion is associated with severe COVID-19 and fecal butyrate level is inversely correlated with pro-inflammatory cytokines IL10 and chemokine CLCX-10.

With its optimized pipeline, using HUMAnN to annotate metagenomic predicted genes are preferable to manually align the predicted genes to databases. According to Franzosa et al. (2018) HUMAnN2 has higher accuracy, sensitivity, and more efficient than other tools like MEGAN (Huson et al., 2011) and COGNIZER (Bose et al., 2015). Furthermore, HUMAnN3 has higher accuracy and true positive rate compared to HUMAnN2 (Beghini et al., 2021).

Enriched urea cycle pathway in COVID-19 patients that are found in Liu et al. (2022) and Zhang et al. (2022) supports Shen et al. (2020) finding that patients with COVID-19 have higher serum concentration of urea. L-ornithine biosynthesis, which also enriched in COVID-19 patients, is very closely related to urea cycle. Several diseases are associated with the dysregulation of ornithine/urea cycle and the enrichment of ornithine, such as infection (i.e., tuberculosis and hepatitis), cancer, and hypertension (Li et al., 2021). However, the mechanism of action behind the association of urea cycle and COVID-19 pathophysiology is still unknown. Several studies suggest that Arginase 1, an enzyme that catalyzes arginine to ornithine and urea, is a metabolic checkpoint in immune response and inflammation and could be activated in immune cells by pro-inflammatory cytokines IL-6 or IL-8 (Li et al., 2021).

In addition, Zhang et al. (2022) link the biological pathway with disease severity. While adjusting age, gender, and comorbidities, they found that sugar derivative degradation, L-isoleucine biosynthesis, and purine nucleotide biosynthesis pathway are negatively correlated with the severity of COVID-19. On the other hand, carbohydrate biosynthesis, purine nucleotide biosynthesis, heme biosynthesis, and peptidoglycan biosynthesis pathway are positively correlated with disease severity.

L-isoleucine is known to induce expression of host defense peptides, such as β-defensins. These peptides help regulate host innate and adaptive immunity and can reduce detrimental effect of pathogens (Gu et al., 2019, Mao et al., 2018). Furthermore, L-isoleucine is inversely correlated with disease severity and CXCL-10, a pro-inflammatory chemokine. L-isoleucine biosynthesis depletion in severe COVID-19 could also be associated with the depletion of *F. prausnitzii*, which is known to produce L-isoleucine (Zhang et al., 2022).

MaAsLin is preferred in differential abundance analysis compared to Kruskal-Wallis H test and Wilcoxon rank-sum test. Besides the lack of data distribution assumption which is not suitable for microbial abundance data, Wilcoxon test is shown to have high FDR rate and relatively low statistical power compared to other method (Lin and Peddada, 2020). We did not find any

study that test the performance of MaAsLin, but Nearing et al. (2022) is including its newest version, MaAsLin2, in their comparative study. MaAsLin2 is shown to have consistent result and high statistical power, while still has a fairly high FDR. MaAsLin2 is also found to perform better with rarefied data.

According to Nearing et al. (2022), LEfSe produces high false discovery rate (FDR) and should be avoided for DA analysis when possible. LEfSe, along with edgeR, is the tool which identified most significant hits that were not identified by any other tool. This high FDR is thought to be resulted from its method of scaling (TSS) and the lack of FDR p-value correction.

A big limitation of this review is to associate the results with the approaches and tools used by the studies. It is very difficult to determine whether the differences in the taxonomic abundance between studies are the effect of different taxonomic classification or differential abundance tools used. The same case applies for differences in functional annotation. Furthermore, it is also very hard to say that the result differences are dependent on the methodology, since various factors also involved in these studies. For instance, these studies differ in their sample size, the periods in which the sample is collected, the comorbidities and the clinical manifestation of the patients, and antibiotic use. Even within each study, these variables could vary. Besides factors involved in the research itself, gut microbiome composition is also affected by other factors unrelated to COVID-19, such as diet, genetic, lifestyles and medication. These studies could produce a very different result when conducted in different geographical region with different diet and genetic make-up.

Furthermore, these studies are just a cross-section observational studies which capture the gut microbiome composition at specific time points. They cannot indicate whether the gut microbiome composition and function variation is determining COVID-19 severity or is it caused by the virus infection itself.

Nevertheless, a lot of benchmarking studies available could give us an idea of the reliability of the approach chosen by these studies. We could then have an informed interpretation about the results of these studies.

In conclusion, Liu et al. (2021) study, which used a distinctly different bioinformatics pipeline with other studies reviewed, does not support or align with the findings in the other studies. The results from each of the four other studies share more similarity with each other but are still quite different. As mentioned above, the results of these studies and microbiota itself is affected by a plethora of other factors and it is very difficult to point out which is responsible for which result. As a researcher, this review could serve as a reminder to be aware and informed of the methods used in metagenomics studies as it might influence the results of the experiment, and to always use the tools that best answer our research questions in our own metagenomics analysis.

References

Abubucker, S., Segata, N., Goll, J., Schubert, A. M., Izard, J., Cantarel, B. L., Rodriguez-Mueller, B., Zucker, J., Thiagarajan, M., Henrissat, B., White, O., Kelley, S. T., Methé, B., Schloss, P. D., Gevers, D., Mitreva, M., & Huttenhower, C. (2012). Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Computational Biology*, *8*(6). https://doi.org/10.1371/journal.pcbi.1002358

Acinas, S. G., Marcelino, L. A., Klepac-Ceraj, V., & Polz, M. F. (2004). Divergence and redundancy of 16S rrna sequences in genomes with multiple *rrn* operons. *Journal of Bacteriology*, *186*(9), 2629–2635. https://doi.org/10.1128/jb.186.9.2629-2635.2004

Alneberg, J., Bjarnason, B. S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., Lahti, L., Loman, N. J., Andersson, A. F., & Quince, C. (2014). Binning metagenomic contigs by coverage and composition. *Nature Methods*, *11*(11), 1144–1146. https://doi.org/10.1038/nmeth.3103

Beghini, F., McIver, L. J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., Mailyan, A., Manghi, P., Scholz, M., Thomas, A. M., Valles-Colomer, M., Weingart, G., Zhang, Y., Zolfo, M., Huttenhower, C., Franzosa, E. A., & Segata, N. (2021). Author response: Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with Biobakery 3. https://doi.org/10.7554/elife.65088.sa2

Bose, T., Haque, M. M., Reddy, C. V. S. K., & Mande, S. S. (2015). Cognizer: A framework for functional annotation of Metagenomic datasets. *PLOS ONE*, *10*(11). https://doi.org/10.1371/journal.pone.0142102

Breyner, N. M., Michon, C., de Sousa, C. S., Vilas Boas, P. B., Chain, F., Azevedo, V. A., Langella, P., & Chatel, J. M. (2017). Microbial anti-inflammatory molecule (MAM) from faecalibacterium prausnitzii shows a protective effect on DNBS and DSS-induced colitis model in mice through inhibition of NF-KB pathway. *Frontiers in Microbiology*, *8*. https://doi.org/10.3389/fmicb.2017.00114

Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C. A., Holland, T. A., Keseler, I. M., Kothari, A., Kubo, A., Krummenacker, M., Latendresse, M., Mueller, L. A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D. S., Weerasinghe, D., Zhang, P., & Karp, P. D. (2013). The METACYC database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic Acids Research*, *42*(D1). https://doi.org/10.1093/nar/gkt1103

Compeau, P. E., Pevzner, P. A., & Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology*, *29*(11), 987–991. https://doi.org/10.1038/nbt.2023

Contini, C., Caselli, E., Martini, F., Maritati, M., Torreggiani, E., Seraceni, S., Vesce, F., Perri, P., Rizzo, L., & Tognon, M. (2020). Covid-19 is a multifaceted challenging pandemic which needs urgent public health interventions. *Microorganisms*, *8*(8), 1228. https://doi.org/10.3390/microorganisms8081228

Darling, A. E., Jospin, G., Lowe, E., Matsen, F. A., Bik, H. M., & Eisen, J. A. (2014). PhyloSift: Phylogenetic Analysis of genomes and metagenomes. *PeerJ*, *2*. https://doi.org/10.7717/peerj.243

Davis-Richardson, A. G., Ardissone, A. N., Dias, R., Simell, V., Leonard, M. T., Kemppainen, K. M., Drew, J. C., Schatz, D., Atkinson, M. A., Kolaczkowski, B., Ilonen, J., Knip, M., Toppari, J., Nurminen, N., HyÃ¶ty, H., Veijola, R., Simell, T., MykkÃ¤nen, J., Simell, O., & Triplett, E. W. (2014). Bacteroides Dorei dominates gut microbiome prior to autoimmunity in Finnish children at high risk for type 1 diabetes. *Frontiers in Microbiology*, *5*. https://doi.org/10.3389/fmicb.2014.00678

Deriu, E., Boxx, G. M., He, X., Pan, C., Benavidez, S. D., Cen, L., Rozengurt, N., Shi, W., & Cheng, G. (2016). Influenza virus affects intestinal microbiota and secondary salmonella infection in the gut through type I interferons. *PLOS Pathogens*, *12*(5). https://doi.org/10.1371/journal.ppat.1005572

Drezen, E., Rizk, G., Chikhi, R., Deltel, C., Lemaitre, C., Peterlongo, P., & Lavenier, D. (2014). GATB: Genome Assembly & Analysis Tool Box. *Bioinformatics*, *30*(20), 2959–2961. https://doi.org/10.1093/bioinformatics/btu406

Dumas, A., Bernard, L., Poquet, Y., Lugo-Villarino, G., & Neyrolles, O. (2018). The role of the lung microbiota and the gut-lung axis in respiratory infectious diseases. *Cellular Microbiology*, *20*(12). https://doi.org/10.1111/cmi.12966

Fernandes, A. D., Reid, J. N. S., Macklaim, J. M., McMurrough, T. A., Edgell, D. R., & Gloor, G. B. (2014). Unifying the analysis of high-throughput sequencing datasets: Characterizing RNA-seq, 16S rrna gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*, *2*(1). https://doi.org/10.1186/2049-2618-2-15

Ferreira-Halder, C. V., Faria, A. V., & Andrade, S. S. (2017). Action and function of faecalibacterium prausnitzii in health and disease. *Best Practice & Research Clinical Gastroenterology*, *31*(6), 643–648. https://doi.org/10.1016/j.bpg.2017.09.011

Franzosa, E. A., McIver, L. J., Rahnavard, G., Thompson, L. R., Schirmer, M., Weingart, G., Lipson, K. S., Knight, R., Caporaso, J. G., Segata, N., & Huttenhower, C. (2018). Species-level functional profiling of metagenomes and metatranscriptomes. *Nature Methods*, *15*(11), 962–968. https://doi.org/10.1038/s41592-018-0176-y

Geirnaert, A., Calatayud, M., Grootaert, C., Laukens, D., Devriese, S., Smagghe, G., De Vos, M., Boon, N., & Van de Wiele, T. (2017). Butyrate-producing bacteria supplemented in vitro to crohn's disease patient microbiota increased butyrate production and enhanced intestinal epithelial barrier integrity. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-11734-8

Gill, S. R., Pop, M., DeBoy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., Gordon, J. I., Relman, D. A., Fraser-Liggett, C. M., & Nelson, K. E. (2006). Metagenomic analysis of the human distal gut microbiome. *Science*, *312*(5778), 1355–1359. https://doi.org/10.1126/science.1124234

Gu, C., Mao, X., Chen, D., Yu, B., & Yang, Q. (2019). Isoleucine plays an important role for maintaining immune function. *Current Protein & Peptide Science*, *20*(7), 644–651. https://doi.org/10.2174/1389203720666190305163135

Harmer, D., Gilbert, M., Borman, R., & Clark, K. L. (2002). Quantitative mrna expression profiling of Ace 2, a novel homologue of angiotensin converting enzyme. *FEBS Letters*, *532*(1-2), 107–110. https://doi.org/10.1016/s0014-5793(02)03640-2

Hofmeyr, S., Egan, R., Georganas, E., Copeland, A. C., Riley, R., Clum, A., Eloe-Fadrosh, E., Roux, S., Goltsman, E., Buluç, A., Rokhsar, D., Oliker, L., & Yelick, K. (2020). Terabase-scale metagenome coassembly with metahipmer. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-67416-5

Huerta-Cepas, J., Forslund, K., Szklarczyk, D., Jensen, L. J., von Mering, C., & Bork, P. (2016). Fast genome-wide functional annotation through orthology assignment by Eggnog-Mapper. https://doi.org/10.1101/076331

A Hussain, S.-R., Santoro, J. L., Rohlfing, M., Salzman, N. H., & Grayson, M. H. (2018). Dysbiosis of intestinal microbiota increases mortality to a respiratory viral infection through elevated production of IFNΓ by innate lymphoid cells. *Journal of Allergy and Clinical Immunology*, *141*(2). https://doi.org/10.1016/j.jaci.2017.12.891

Kalantar-Zadeh, K., Ward, S. A., Kalantar-Zadeh, K., & El-Omar, E. M. (2020). Considering the effects of microbiome and diet on SARS-COV-2 infection: Nanotechnology roles. *ACS Nano*, *14*(5), 5179–5182. https://doi.org/10.1021/acsnano.0c03402

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2013). Data, information, knowledge and principle: Back to metabolism in kegg. *Nucleic Acids Research*, *42*(D1). https://doi.org/10.1093/nar/gkt1076

Kang, D. D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., & Wang, Z. (2019). MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from Metagenome Assemblies. *PeerJ*, *7*. https://doi.org/10.7717/peerj.7359

Keely, S., Talley, N. J., & Hansbro, P. M. (2011). Pulmonary-intestinal cross-talk in mucosal inflammatory disease. *Mucosal Immunology*, *5*(1), 7–18. https://doi.org/10.1038/mi.2011.55

Kumar, M. S., Slud, E. V., Okrah, K., Hicks, S. C., Hannenhalli, S., & Corrada Bravo, H. (2018). Analysis and correction of compositional bias in sparse sequencing count data. *BMC Genomics*, *19*(1). https://doi.org/10.1186/s12864-018-5160-5

Lake, M. A. (2020). What we know so far: Covid-19 current clinical knowledge and research. *Clinical Medicine*, *20*(2), 124–127. https://doi.org/10.7861/clinmed.2019-coron

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, *15*(2). https://doi.org/10.1186/gb-2014-15-2-r29

Lee, W.-P., Stromberg, M. P., Ward, A., Stewart, C., Garrison, E. P., & Marth, G. T. (2014). Mosaik: A hash-based algorithm for accurate next-generation sequencing short-read mapping. *PLoS ONE*, *9*(3). https://doi.org/10.1371/journal.pone.0090581

Leung, W. K., To, K.-fai, Chan, P. K. S., Chan, H. L. Y., Wu, A. K. L., Lee, N., Yuen, K. Y., & Sung, J. J. Y. (2003). Enteric involvement of severe acute respiratory syndrome-associated coronavirus infection. *Gastroenterology*, *125*(4), 1011–1017. https://doi.org/10.1016/s0016-5085(03)01215-0

Li, D., Liu, C.-M., Luo, R., Sadakane, K., & Lam, T.-W. (2015). Megahit: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, *31*(10), 1674–1676. https://doi.org/10.1093/bioinformatics/btv033

Li, T., Ning, N., Li, B., Luo, D., Qin, E., Yu, W., Wang, J., Yang, G., Nan, N., He, Z., Yang, N., Gong, S., Li, J., Liu, A., Sun, Y., Li, Z., Jia, T., Gao, J., Zhang, W., … Wang, H. (2021). Longitudinal metabolomics reveals ornithine cycle dysregulation correlates with inflammation and coagulation in COVID-19 severe patients. *Frontiers in Microbiology*, *12*. https://doi.org/10.3389/fmicb.2021.723818

Lin, H., & Peddada, S. D. (2020). Analysis of microbial compositions: A review of normalization and differential abundance analysis. *Npj Biofilms and Microbiomes*, *6*(1). https://doi.org/10.1038/s41522-020-00160-w

Liu, B., Gibbons, T., Ghodsi, M., Treangen, T., & Pop, M. (2011). Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. *Genome Biology*, *12*(S1). https://doi.org/10.1186/1465-6906-12-s1-p11

Liu, Q., Mak, J. W., Su, Q., Yeoh, Y. K., Lui, G. C.-Y., Ng, S. S., Zhang, F., Li, A. Y., Lu, W., Hui, D. S.-C., Chan, P. K. S., Chan, F. K., & Ng, S. C. (2022). Gut microbiota dynamics in a prospective cohort of patients with post-acute COVID-19 syndrome. *Gut*, *71*(3), 544–552. https://doi.org/10.1136/gutjnl-2021-325989

Liu, Y., Zhang, H., Tang, X., Jiang, X., Yan, X., Liu, X., Gong, J., Mew, K., Sun, H., Chen, X., Zou, Z., Chen, C., & Qiu, J. (2021). Distinct metagenomic signatures in the SARS-COV-2 infection. *Frontiers in Cellular and Infection Microbiology*, *11*. https://doi.org/10.3389/fcimb.2021.706970

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with deseq2. *Genome Biology*, *15*(12). https://doi.org/10.1186/s13059-014-0550-8

Ma, W.-T., Pang, M., Fan, Q.-L., & Hua, J.-L. (2019). The commensal microbiota and viral infection: A comprehensive review. *Frontiers in Immunology*, *10*. https://doi.org/10.3389/fimmu.2019.01551

Madeira, F., Pearce, M., Tivey, A. R., Basutkar, P., Lee, J., Edbali, O., Madhusoodanan, N., Kolesnikov, A., & Lopez, R. (2022). Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Research*, *50*(W1). https://doi.org/10.1093/nar/gkac240

Mao, X., Gu, C., Ren, M., Chen, D., Yu, B., He, J., Yu, J., Zheng, P., Luo, J., Luo, Y., Wang, J., Tian, G., & Yang, Q. (2018). L-isoleucine administration alleviates rotavirus infection and immune response in the weaned piglet model. *Frontiers in Immunology*, *9*. https://doi.org/10.3389/fimmu.2018.01654

McHardy, A. C., Martín, H. G., Tsirigos, A., Hugenholtz, P., & Rigoutsos, I. (2006). Accurate phylogenetic classification of variable-length DNA fragments. *Nature Methods*, *4*(1), 63–72. https://doi.org/10.1038/nmeth976

McLachlan, G. J. (2004). *Discriminant analysis and Statistical Pattern Recognition*. John Wiley & Sons.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L., Tosatto, S. C., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2020). Pfam: The Protein Families Database in 2021. *Nucleic Acids Research*, *49*(D1). https://doi.org/10.1093/nar/gkaa913

Morgan, X. C., Kabakchiev, B., Waldron, L., Tyler, A. D., Tickle, T. L., Milgrom, R., Stempak, J. M., Gevers, D., Xavier, R. J., Silverberg, M. S., & Huttenhower, C. (2015). Associations between host gene expression, the mucosal microbiome, and clinical outcome in the pelvic pouch of patients with inflammatory bowel disease. *Genome Biology*, *16*(1). https://doi.org/10.1186/s13059-015-0637-x

Morgan, X. C., Tickle, T. L., Sokol, H., Gevers, D., Devaney, K. L., Ward, D. V., Reyes, J. A., Shah, S. A., LeLeiko, N., Snapper, S. B., Bousvaros, A., Korzenik, J., Sands, B. E., Xavier, R. J., & Huttenhower, C. (2012). Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biology*, *13*(9). https://doi.org/10.1186/gb-2012-13-9-r79

Nearing, J. T., Douglas, G. M., Hayes, M. G., MacDonald, J., Desai, D. K., Allward, N., Jones, C. M., Wright, R. J., Dhanani, A. S., Comeau, A. M., & Langille, M. G. (2022). Microbiome differential abundance methods produce different results across 38 datasets. *Nature Communications*, *13*(1). https://doi.org/10.1038/s41467-022-28034-z

Noguchi, H., Park, J., & Takagi, T. (2006). Metagene: Prokaryotic gene finding from environmental Genome Shotgun sequences. *Nucleic Acids Research*, *34*(19), 5623–5630. https://doi.org/10.1093/nar/gkl723

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). Metaspades: A new versatile metagenomic assembler. *Genome Research*, *27*(5), 824–834. https://doi.org/10.1101/gr.213959.116

Paulson, J. N., Stine, O. C., Bravo, H. C., & Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nature Methods*, *10*(12), 1200–1202. https://doi.org/10.1038/nmeth.2658

Polz, M. F., & Cavanaugh, C. M. (1998). Bias in template-to-product ratios in multitemplate PCR. *Applied and Environmental Microbiology*, *64*(10), 3724–3730. https://doi.org/10.1128/aem.64.10.3724-3730.1998

Ren, Y.-D., Ye, Z.-S., Yang, L.-Z., Jin, L.-X., Wei, W.-J., Deng, Y.-Y., Chen, X.-X., Xiao, C.-X., Yu, X.-F., Xu, H.-Z., Xu, L.-Z., Tang, Y.-N., Zhou, F., Wang, X.-L., Chen, M.-Y., Chen, L.-G., Hong, M.-Z., Ren, J.-L., & Pan, J.-S. (2017). Fecal microbiota transplantation induces hepatitis B virus e-antigen (HBEAG) clearance in patients with positive hbeag after long-term antiviral therapy. *Hepatology*, *65*(5), 1765–1768. https://doi.org/10.1002/hep.29008

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and Microarray Studies. *Nucleic Acids Research*, *43*(7). https://doi.org/10.1093/nar/gkv007

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-Seq Data. *Genome Biology*, *11*(3). https://doi.org/10.1186/gb-2010-11-3-r25

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., & Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, *9*(8), 811–814. https://doi.org/10.1038/nmeth.2066

Sharpton, T. J. (2014). An introduction to the analysis of shotgun metagenomic data. *Frontiers in Plant Science*, *5*. https://doi.org/10.3389/fpls.2014.00209

Shen, B., Yi, X., Sun, Y., Bi, X., Du, J., Zhang, C., Quan, S., Zhang, F., Sun, R., Qian, L., Ge, W., Liu, W., Liang, S., Chen, H., Zhang, Y., Li, J., Xu, J., He, Z., Chen, B., … Guo, T. (2020). Proteomic and metabolomic characterization of COVID-19 patient Sera. *Cell*, *182*(1). https://doi.org/10.1016/j.cell.2020.05.032

Song, L., Huang, Y., Liu, G., Li, X., Xiao, Y., Liu, C., Zhang, Y., Li, J., Xu, J., Lu, S., & Ren, Z. (2022). A novel immunobiotics bacteroides dorei ameliorates influenza virus infection in mice. *Frontiers in Immunology*, *12*. https://doi.org/10.3389/fimmu.2021.828887

Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R., & Wu, C. H. (2007). UNIREF: Comprehensive and non-redundant Uniprot Reference Clusters. *Bioinformatics*, *23*(10), 1282–1288. https://doi.org/10.1093/bioinformatics/btm098

Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., & Segata, N. (2015). Erratum: Metaphlan2 for enhanced metagenomic taxonomic profiling. *Nature Methods*, *13*(1), 101–101. https://doi.org/10.1038/nmeth0116-101b

Villanueva-Millán, M. J., Pérez-Matute, P., & Oteo, J. A. (2015). Gut microbiota: A key player in health and disease. A review focused on obesity. *Journal of Physiology and Biochemistry*, *71*(3), 509–525. https://doi.org/10.1007/s13105-015-0390-3

Vázquez-Castellanos, J. F., Serrano-Villar, S., Jiménez-Hernández, N., Soto del Rio, M. D., Gayo, S., Rojo, D., Ferrer, M., Barbas, C., Moreno, S., Estrada, V., Rattei, T., Latorre, A., Moya, A., & Gosalbes, M. J. (2018). Interplay between gut microbiota metabolism and inflammation in HIV infection. *The ISME Journal*, *12*(8), 1964–1976. https://doi.org/10.1038/s41396-018-0151-8

Wang, H., Wang, H., Sun, Y., Ren, Z., Zhu, W., Li, A., & Cui, G. (2021). Potential associations between microbiome and covid-19. *Frontiers in Medicine*, *8*. https://doi.org/10.3389/fmed.2021.785496

Wang, W.-L., Xu, S.-Y., Ren, Z.-G., Tao, L., Jiang, J.-W., & Zheng, S.-S. (2015). Application of metagenomics in the human gut microbiome. *World Journal of Gastroenterology*, *21*(3), 803. https://doi.org/10.3748/wjg.v21.i3.803

Warton, D. I., & Hui, F. K. (2011). The arcsine is asinine: The analysis of proportions in ecology. *Ecology*, *92*(1), 3–10. https://doi.org/10.1890/10-0340.1

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J. R., Vázquez-Baeza, Y., Birmingham, A., Hyde, E. R., & Knight, R. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, *5*(1). https://doi.org/10.1186/s40168-017-0237-y

Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcostsdata. Accessed 26 July 2022.

Wiersinga, W. J., Rhodes, A., Cheng, A. C., Peacock, S. J., & Prescott, H. C. (2020). Pathophysiology, transmission, diagnosis, and treatment of coronavirus disease 2019 (covid-19). *JAMA*, *324*(8), 782. https://doi.org/10.1001/jama.2020.12839

Wu, J., Liu, J., Zhao, X., Liu, C., Wang, W., Wang, D., Xu, W., Zhang, C., Yu, J., Jiang, B., Cao, H., & Li, L. (2020). Clinical characteristics of imported cases of coronavirus disease 2019 (covid-19) in Jiangsu Province: A multicenter descriptive study. *Clinical Infectious Diseases*, *71*(15), 706–712. https://doi.org/10.1093/cid/ciaa199

Wu, M., & Scott, A. J. (2012). Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics*, *28*(7), 1033–1034. https://doi.org/10.1093/bioinformatics/bts079

Wu, Y.-W., Simmons, B. A., & Singer, S. W. (2015). MaxBin 2.0: An automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, *32*(4), 605–607. https://doi.org/10.1093/bioinformatics/btv638

Xu, J., Liang, R., Zhang, W., Tian, K., Li, J., Chen, X., Yu, T., & Chen, Q. (2020). *faecalibacterium prausnitzii* -derived microbial anti-inflammatory molecule regulates intestinal integrity in diabetes mellitus mice via modulating tight junction protein expression. *Journal of Diabetes*, *12*(3), 224–236. https://doi.org/10.1111/1753-0407.12986

Yao, Y., Cai, X., Fei, W., Ye, Y., Zhao, M., & Zheng, C. (2020). The role of short-chain fatty acids in immunity, inflammation and metabolism. *Critical Reviews in Food Science and Nutrition*, *62*(1), 1–12. https://doi.org/10.1080/10408398.2020.1854675

Ye, S. H., Siddle, K. J., Park, D. J., & Sabeti, P. C. (2019). Benchmarking metagenomics tools for taxonomic classification. *Cell*, *178*(4), 779–794. https://doi.org/10.1016/j.cell.2019.07.010

Ye, Y., & Doak, T. G. (2009). A parsimony approach to biological pathway reconstruction/inference for genomes and Metagenomes. *PLoS Computational Biology*, *5*(8). https://doi.org/10.1371/journal.pcbi.1000465

Yeoh, Y. K., Zuo, T., Lui, G. C.-Y., Zhang, F., Liu, Q., Li, A. Y. L., Chung, A. C. K., Cheung, C. P., Tso, E. Y. K., Fung, K. S. C., Chan, V., Ling, L., Joynt, G., Hui, D. S.-C., Chow, K. M., Ng, S. S., Li, T. C.-M., Ng, R. W. Y., Yip, T. C. F., … Ng, S. C. (2021). Gut microbiota composition reflects disease severity and dysfunctional immune responses in patients with COVID-19. *Gut*, *70*(4), 698–706. https://doi.org/10.1136/gutjnl-2020-323020

Yildiz, S., Mazel-Sanchez, B., Kandasamy, M., Manicassamy, B., & Schmolke, M. (2018). Influenza A virus infection impacts systemic microbiota dynamics and causes quantitative enteric dysbiosis. *Microbiome*, *6*(1). https://doi.org/10.1186/s40168-017-0386-z

Yoshida, N., Emoto, T., Yamashita, T., Watanabe, H., Hayashi, T., Tabata, T., Hoshi, N., Hatano, N., Ozawa, G., Sasaki, N., Mizoguchi, T., Amin, H. Z., Hirota, Y., Ogawa, W., Yamada, T., & Hirata, K.-ichi. (2018). *bacteroides vulgatus* and *bacteroides dorei* reduce gut microbial lipopolysaccharide production and inhibit atherosclerosis. *Circulation*, *138*(22), 2486–2498. https://doi.org/10.1161/circulationaha.118.033714

Zhang, F., Wan, Y., Zuo, T., Yeoh, Y. K., Liu, Q., Zhang, L., Zhan, H., Lu, W., Xu, W., Lui, G. C. Y., Li, A. Y. L., Cheung, C. P., Wong, C. K., Chan, P. K. S., Chan, F. K. L., & Ng, S. C. (2022). Prolonged impairment of short-chain fatty acid and L-isoleucine biosynthesis in gut microbiome in

patients with covid-19. *Gastroenterology*, *162*(2). https://doi.org/10.1053/j.gastro.2021.10.013

Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., Si, H.-R., Zhu, Y., Li, B., Huang, C.-L., Chen, H.-D., Chen, J., Luo, Y., Guo, H., Jiang, R.-D., Liu, M.-Q., Chen, Y., Shen, X.-R., Wang, X., … Shi, Z.-L. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, *579*(7798), 270–273. https://doi.org/10.1038/s41586-020-2012-7

Zhou, Y., Hou, Y., Shen, J., Huang, Y., Martin, W., & Cheng, F. (2020). Network-based drug repurposing for Novel Coronavirus 2019-ncov/SARS-COV-2. *Cell Discovery*, *6*(1). https://doi.org/10.1038/s41421-020-0153-3

Zuo, T., Zhang, F., Lui, G. C. Y., Yeoh, Y. K., Li, A. Y. L., Zhan, H., Wan, Y., Chung, A. C. K., Cheung, C. P., Chen, N., Lai, C. K. C., Chen, Z., Tso, E. Y. K., Fung, K. S. C., Chan, V., Ling, L., Joynt, G., Hui, D. S. C., Chan, F. K. L., … Ng, S. C. (2020). Alterations in gut microbiota of patients with covid-19 during time of hospitalization. *Gastroenterology*, *159*(3). https://doi.org/10.1053/j.gastro.2020.05.048