Unlocking the diagnostic and therapeutic potential of metagenomics

Caitlin Loeffler¹, Keylie M. Gibson², Lana Martin³, Liz Chang⁴, Jeremy Rotman⁵, Ian V. Toma⁶, Christopher E. Mason⁷,², Eleazar Eskin¹,⁵, Joseph P. Zackular⁹, Keith A. Crandall¹⁰, David Koslicki¹¹,¹²*, Serghei Mangul⁴*  

*These authors contributed equally to this work

¹Department of Computational Medicine, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA  
²Computational Biology Institute, George Washington University, Washington, DC, USA  
³Department of Clinical Pharmacy, School of Pharmacy, University of Southern California, Los Angeles, CA, USA  
⁴Department of Clinical Pharmacy, School of Pharmacy, University of Southern California, Los Angeles, CA, USA  
⁵Department of Computer Science, University of California Los Angeles, Los Angeles, CA, USA  
⁶Clinical Research & Leadership, School of Medicine and Health Sciences, George Washington University, Washington, DC, USA  
⁷Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, USA
The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, USA

8The WorldQuant Initiative for Quantitative Prediction, Weill Cornell Medicine, New York, NY, USA

9Department of Pathology and Laboratory Medicine, University of Pennsylvania and Children’s Hospital of Philadelphia, Philadelphia, PA, USA

10Computational Biology Institute, George Washington University, Washington, DC, USA
Department of Biostatistics & Bioinformatics, George Washington University, Washington, DC, USA

11Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA, USA

12Department of Biology, The Pennsylvania State University, University Park, PA, USA
Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA, USA
**Abstract**

During the past decade, rapid advancements in sequencing technologies have enabled the study of human-associated microbiomes at an unprecedented scale. Metagenomics, specifically, is emerging as a clinical tool to identify the agents of infection, track the spread of diseases, and surveil potential pathogens. Yet, despite advances in high-throughput sequencing technologies and bioinformatics algorithms, metagenomics still has limitations barring mass clinical acceptance and broad implementation. Algorithms currently struggle to balance sensitivity and specificity of discovered biomarkers, genomic reference databases are incomplete, and studies are frequently not reproducible beyond the original environmental settings. Once these hurdles are overcome, clinical metagenomics will be able to inform doctors of the best, targeted treatment for their patients and provide early detection of disease. Here we present an overview of metagenomics methods with a discussion of computational challenges and limitations. Our review covers clinical applications and presents applications for strain identification of *Mycobacterium tuberculosis* (TB), identification of infectious agents in cases of encephalitis, biomarkers for Inflammatory Bowel Disease, and explores the challenges of sampling the microbiome in various tissues. Our review also summarizes 11 of the most recent clinical metagenomics studies and discusses their commonalities and unique features.
Introduction

During the past decade, rapid advancements in genomics technologies have enabled the study of human-associated microbiomes at an unprecedented scale. Together with a significant reduction in cost of sequencing technologies and development of advanced gene-editing tools, the potential of leveraging human-associated microbiome data for development of novel diagnostic and therapeutic tools is enormous. Medical applications include, but are not limited to, infection detection, surveillance and tracking of outbreaks, antimicrobial resistance prediction and strain resolution, and disease prognosis. However, integration of the field of metagenomics with clinical medicine has yet to be accomplished (Figure 1a). A large body of work has been published in the field of metagenomics, while papers addressing clinical applications of metagenomics are just starting to appear (Figure 1b).

A microbiome is the microscopic ecosystem of a given habitat, which includes both biotic and abiotic elements. Microbiota are the communities of microorganisms (bacteria, fungi, protists, parasites, viruses) that simultaneously exist and interact within a microbiome\(^1\). All environments on Earth, and even the space station above Earth, include dynamic microbial communities that live on, inside, and around host microbiota. Two methodologies for identifying and analyzing the microbiome dominate today’s microbiota studies: culture-dependent and culture-independent methods.
Figure 1. (a) There is a gap between metagenomics research and application of metagenomics discoveries to clinical developments. The missing pieces are standardization of methods and data, consensus on best tools and practices, and clinical validation. (b) The cumulative trend of
papers published on metagenomics (blue) and clinical metagenomics (orange). Note: the vertical axis is log-based to demonstrate cumulation of papers published in PubMed since 2004.
Figure 2. (a) Sample collection. The first step of clinical metagenomics is sample collection from the patient. Samples take the form of any fluid (e.g., blood, urine), solid (e.g., feces), or alternate sampling method (e.g., swabs) taken from the human body. (b) Host-DNA depletion
or microbial enrichment. Samples taken from living hosts may contain an over abundance of host DNA that may overshadow reads; often, host DNA can be mostly removed through host-DNA depletion although at a financial cost as well as loss of some microbial material. Alternatively microbial enrichment will increase the abundance of microbial DNA in the sample. Enrichment occurs after DNA extraction, where as host DNA depletion occurs before. These steps are optional. (c) DNA sequencing. The DNA within the samples is extracted and sequenced. Five sequencing platforms are Illumina, ThermoFisher (not shown), BGI (not shown), Nanopore, and PacBio. The latter two generate long-reads (<1,000 bp) while the former three generate paired-end short-reads (>300bp). Illumina, BGI, and PacBio determine base pairs utilizing fluorescently tagged nucleotides and *taq* Polymerase. Nanopore determines base pairs by passing a single stranded DNA through a nanopore, the unique interruption in the ion current through the nanopore differs by the set of nucleotides in the pore, and ThermoFisher also uses pH changes as bases are incorporated to sequence. A computer is capable of guessing (with a 13% error rate) what nucleotides are in the pore. (d) Sequencing technology statistics. Illumina has the lowest single-pass sequencing error rate, but the shortest read length. Illumina determines base pairs by clustering clonally-amplified copies and selecting the dominant color. During the sequencing process, strands lag behind or pull ahead of the current base, and the signal becomes obscured. This affects accuracy at a distance close to 300bp. PacBio, which produces the longest read lengths of the three platforms, does not rely on clusters of identical strands remaining in step, but can instead sequence a single strand of DNA one time (10% error rate) or multiple times (<0.1% error rate from circular consensus, a.k.a. HiFi reads). This single-molecule, multiple pass technique gives PacBio the lowest error rate of the three technologies, and the
errors are randomly distributed throughout the read. Nanopore has the longest read length and the highest error rate. The error rate is due to sequencing read error. When reading bases through the pores, there are multiple nucleotides in the channel at one time which can obscure the base pair identification, leading to the high error rate. Illumina had the highest throughput followed by PacBio and Nanopore. (Illumina statistics based on NovaSeq6000, nanopore based on PromethION, PacBio estimated based on Sequel II) (e) **Bioinformatics analysis.** There are three types of mapping algorithms used to identify reads and measure organism abundance. The first is mapping-based profiling which maps the entire read to the entire reference. The second is marker based profiling, which only maps the reads to parts of the genome that are unique to each species or strain (red and orange regions) and not to regions that are homologous (green and blue). This step reduces the time it takes to map the reads. Finally, k-mer based profiling, which breaks up the reads and references into substrings and compares the substrings. (Binning methods are not discussed.) (f) **Clinical applications.** After the reads identities and abundances have been found, Bioinformatics analysis returns a report that can inform clinical decisions. Clinical metagenomics can be used to detect infections, track the progress of outbreaks, predict antimicrobial resistance, and inform doctors of the best disease prognosis.

Culture-dependent methods (CDMs) refer to any specific method that requires the growth of sample microbes in an artificial environment (i.e., culture) in order to identify, quantify, and phenotypically characterize microbial organisms. These methods are currently the gold standard for the classification of microbes isolated from their host environment, including those from the
human body. CDMs are widely applied in clinical and environmental laboratories in order to conduct phenotypic discrimination of bacterial colonies, examination of intracellular pathogenicity, and susceptibility to antibiotics. The application of Next-Generation Sequencing (NGS) technologies, and subsequent sequence analyses that use specialized computational methods, to CDMs has allowed for easier and more refined classification of microbes. CDMs are capable of generating accurate inventories of microbial taxa that can grow on the selected media and have demonstrated utility in clinical diagnosis.

Culture-independent methods (CIMs) include all methods that do not rely on the growth of microbes in culture (e.g., targeted amplicon sequencing, metagenomics, etc.). CIMs can deliver a complete and unbiased report on the microbes present in a sample. Advancements in high throughput technologies have led to a wide applicability of CIMs in microbiome research and medicine. Such techniques are used to investigate the genetic material of human or environmentally-associated microbial samples. Since CIMs do not need to allocate time for bacterial growth, identification of strains and organisms can take place in less time and with higher accuracy than with CDMs. In addition, CIMs are capable of isolating novel genomes and identifying microorganisms that cannot be cultured.

CIMs are currently the microbial-identifying methods that offer the most flexibility and are capable of delivering the most comprehensive results. Our review discusses the advantages, limitations, and potential clinical applications of one class of CIMs: metagenomics methods.
Current Approaches to Metagenomics

The results of metagenomic studies shed light on the biotic composition of the microbiome and the function of the microbial organisms’ genes, both of which can inform translational developments in the fields of biology, public health, and medicine (Table 1). A metagenomics study considers the genes and genomes of all microorganisms from a microbiome (i.e., the metagenome). The DNA composing the metagenome is collected directly from a sample of the studied host environment (Figure 2a), sheared, and sequenced in order to isolate the base pair patterns (Figure 2c). This process, called shotgun sequencing, produces a number of sequences (referred to as reads) that can be assembled into larger contiguous pieces and taxonomically characterized (Figure 2e). Generating and analyzing results requires the utilization of complex computational algorithms.
Table 1. Clinical applications and obstacles of applying metagenomic NGS (mNGS). Clinical applications of metagenomics methods are a few of the specific areas of clinics to which metagenomics methods can be applied. Requirement for metagenomics analysis covers the features of methods that are needed for metagenomics to be validated within the given clinical application. Sequencing technologies list the current NGS technologies that would be needed for the given clinical application. Illumina and IonTorrent represent short-read sequencers. PacBio and Nanopore represent long-read sequencers. 16S, 18S, and ITS represents amplicon sequencing. Amplicon sequencing is not metagenomics, but rather metataxonomy.

| Clinical Application                          | Requirement for metagenomics analysis                  | Sequencing Technology                  | Databases                      |
|----------------------------------------------|-------------------------------------------------------|----------------------------------------|--------------------------------|
| Infection detection                          | Highly-accurate methods                                | Illumina/IonTorrent                    | Complete                       |
|                                              |                                                       | PacBio/Nanopore                         |                                |
|                                              |                                                       | 16S/18S/ITS                            |                                |
| Surveillance and tracking of outbreaks       | Sample collection across large number of individuals  | Illumina/IonTorrent                    | Sequencing Technology Specific |
|                                              |                                                       | 16S/18S/ITS                            |                                |
| Rapid diagnostics                            | Speed (<4h)                                           | Nanopore/PacBio                         | Sequencing Technology Specific |
| Antimicrobial resistance prediction, strain resolution | Low read error rate                                | PacBio/Nanopore                         | Custom                         |
| Disease prognosis                            | Modeling dynamics                                     | Illumina/IonTorrent                    | Custom                         |
|                                              |                                                       | PacBio/Nanopore                         |                                |
|                                              |                                                       | 16S/18S/ITS                            |                                |
Today’s computational algorithms for metagenomic analysis have limitations. Genetic data obtained through metagenomics methods contain an increased number of DNA fragments that must be aligned against hundreds of thousands of microbial genomes in order to be analytically useful. This increased number of DNA fragments — and the large size of the reference database — poses unique challenges to existing metagenomic methods and the computational resources they depend on. Additional challenges include the lack of comprehensive gene catalogs, biases in functional profiling, and lack of standardization in publishing raw data. Difficulty integrating meta-omics analysis tools with existing frameworks also limit and challenge the potential use of metagenomic methods.

By identifying microbes from diverse sample types and determining their impact on the microbiome, metagenomics has substantial potential as a tool for human health and the diagnosis and treatment of disease. Human biomedical studies focus on the variety of microbial diversity across body compartments (e.g., gut, skin, nasal, oral) as well as niche specialization within those compartments and how these variables influence the bodies reaction to treatments. For example, a comparison of gut microbiomes shows that varied communities of microbiota can influence an individual’s reactions to drugs and treatments. Metagenomic techniques allow for the capture of genetic features (e.g., functional elements, protein domains, or variants) of microbes present in the sample, which are then compared to other samples across health and disease status. As costs continue to decrease and studies begin to elucidate the therapeutic potential of microbiome characterizations, there is a strong potential for the integration of these
tools in the clinic. The application of metagenomics in clinical metagenomics has yet to be fully explored as clinical metagenomics studies have begun to be published only recently (Figure 1b).

Our review of the rapidly evolving approaches to metagenomics explains the limitations of CDMs and the rise of CIMs, as well as targeted and untargeted sequencing methods. We outline the complexities and limitations of computational algorithms that handle output from sequencing technologies. Finally, we provide examples of metagenomic methods, which show promise as a valuable tool for developing translational discoveries in clinical laboratories.

The rise of culture-independent genomic-based microbiome profiling

CDMs are the gold standard for microbe identification and analysis in pathology research and practice. Microbe morphology, growth, and response to environmental variables (e.g., antibiotics, gel composition) have been used for almost 300 years to study and identify microbes⁸. CDMs create artificial, controlled environments using a nutrient-rich media which can be designed to enrich diverse groups of inure organisms or grow microbes with specific or unknown requirements. Once cultivated, these organisms can be phenotypically characterized using different microbiological and biochemical diagnosis techniques.

The Sanger automated sequencing method⁹ was developed in the mid-1970s and allowed researchers, for the first time, to sequence the base pairs in segments of DNA (called reads). Once the content of DNA could be analyzed, researchers started viewing the base pair pattern
within the DNA as molecular biomarkers for the classification of microorganisms\textsuperscript{10}. At the time, the Sanger method was an advancement on CDMs and allowed pathologists to more accurately identify microbes from culture.

The 16S rRNA gene is the dominant marker in microbial studies and exemplifies the revolutionary nature of Sanger sequencing. This region of the genome codes for a segment of the small ribosomal subunit, called the Shine-Dalgarno sequence\textsuperscript{11}, which is required for the proper initiation, elongation, and termination of DNA translation in prokaryotes\textsuperscript{12} (the homologous region in eukaryotic cells is the 18S rRNA gene\textsuperscript{13}). Since this segment is vitally important for the development and survival of the microbe, it contains regions with very low mutation rates, making it ideal for use as a molecular classification marker at higher taxonomic levels (e.g. kingdom, phylum)\textsuperscript{10,14}. The 16S region also contains regions that are hypervariable\textsuperscript{15}, meaning the regions have high mutation rates. Thus, the 16S region is ideal for use as a molecular classification marker at the lower taxonomic levels (e.g., genus, species)\textsuperscript{14,16–18}. Indeed, the ubiquitous use of this single genetic marker to identify microbes has led to the formation of extensive reference databases just for this gene (e.g., Greengenes\textsuperscript{19}, Silva\textsuperscript{20}).
Despite advancements in sequencing technologies, multiple factors limit the application of CDMs to clinical translation efforts. Culturing procedures are slow, labor intensive, and cannot be scaled across a large number of samples. Additionally, profiling beyond strains with CDMs alone is challenging. Moreover, much of the microbial diversity observed in the microbiome have yet to be cultured due to microbiologists being unable to replicate biologically crucial aspects of their host environment in the laboratory setting. Microbiologists do not yet know enough about the nutrient and environmental requirements of many microbes to predict what is necessary for these organisms to grow. Furthermore, applying such techniques to potentially life-threatening bacteria poses safety concerns, and CDMs have limited potential to detect viruses, which usually cannot be cultured without host cells. These limitations affected the applicability of Sanger sequencing during the 1970s because this type of sequencing requires high concentrations of pure DNA that are not always available — especially if the target microbe
cannot be cultured\textsuperscript{23}. For the study of uncultured microbes, reliance on CDMs was a substantial hurdle to metagenomics and clinical research.

CIMs first became possible in the late 1980s with Kary Mullis’ development of Polymerase Chain Reaction (PCR)\textsuperscript{24}. PCR artificially amplifies specific regions of DNA in order to deepen coverage of those regions for sequencing, functional analysis of a specific gene, or detection of polymorphisms or point mutations\textsuperscript{24–26}. Among other applications, PCR is capable of amplifying the 16S rRNA gene of microbes, which isolates the mutation-informative region of DNA for sequencing. This can be applied to both cultured microbes (to isolate the 16S region for sequencing and analysis) or uncultured microbes. The latter reveals the true value of PCR to the microbiology field: it allows for the study and identification of microbes that researchers had not yet been able to grow in culture, even if the amount of DNA available at the outset was incredibly small. PCR also allowed clinics to diagnose infectious pathogens, such as \textit{Mycobacterium leprae} (a bacterium causing leprosy)\textsuperscript{27–29}, which researchers and pathologists had continually failed to grow and classify using CDMs\textsuperscript{30}. 
**Figure 4.** (a) Clinical applications and the required features of metagenomics methods.

Differences in (b) long read and (c) short read sequencing completeness given the same cost per sample. Long read sequencing obtains reads from few species, but read coverage is high — representing almost the complete coverage of the species genome. Short read sequencing obtains reads from many species, but reads only represent a small fraction of the genome — representing only a small portion of the assembly of the species genome. Read coverage is represented in blue.

Despite the emergence of PCR, during the 80s and 90s, the study of the entire genetic content of all the organisms in the microbiome (called the metagenome) was limited. The cost of using the Sanger sequencing method presented a serious challenge. Beyond the expense, reliance on PCR to amplify the DNA segments of organisms that are too low in abundance for sequencing poses several challenges. Utilizing PCR before sequencing introduces additional mutations which
decreases the accuracy of the results\textsuperscript{31}. Moreover, PCR requires the presence of an artificially constructed primer that binds downstream of the target gene, thus allowing \textit{taq} DNA polymerase to bind the strand and begin replication\textsuperscript{25,32}. Therefore, the base pair pattern that occurs before the target sequence (the binding site of the primer) must already be known in order for PCR to work\textsuperscript{25}. This leads to an increased abundance of DNA that bind well with the primer, which may not reflect the associated species actual abundance in the sample. It was not until the development of cheaper sequencing technologies (termed Next Generation Sequencing) in the 2000s that the full metagenome could be studied.

Today, Sanger is far from the only available sequencing method. Sequencing is increasingly done using Next Generation Sequencing (NGS) technologies, which provide researchers with a dramatically increased throughput (\textbf{Figure 3c}; \textbf{Figure 2e}). When compared to prior sequencing methods, NGS technologies yield reads at a rate that is orders of magnitude higher per run and at a fraction of the cost per base\textsuperscript{8,33} (\textbf{Figure 3c}; \textbf{Supplementary Note 1}). Today, Illumina leads the market in sequencing technologies that use short-reads (<300 bp) and carry the lowest cost per Gigabase pair (Gbp) (\textbf{Figure 3a}; \textbf{Supplementary Note 1}). Pacific Bioscience (PacBio) and Oxford Nanopore Technologies have developed platforms capable of sequencing long-reads (>10,000 - 2.2M bp). While these long-read platforms cost more and have a higher error rate when compared to Illumina’s short-reads\textsuperscript{8} (\textbf{Figure 3b}; \textbf{Figure 3d}), they are still an improvement on Sanger with drastically lower costs per Gbp (\textbf{Supplementary Note 1}). In terms of throughput, long-read technologies have yet to match Illumina (\textbf{Figure 3e}). Short-reads are capable of covering more microbes than long-reads, however, these reads are capable of covering less area
of the genome than long-reads can (Figure 4b; Figure 4c). Bacteriophages and viruses have genomes sized between a few thousand bases, where long-read sequencing technologies can cover the entirety of the genome, to several hundred kilobases. Bacterial genomes are 0.5 Mb to 10 Mb in length. Eukaryotic genome length varies across and within kingdoms from approximately 10 Mb in some fungi to more than 100000 Mb in certain plants.\textsuperscript{34}

These advances in NGS technologies have allowed for the creation of metagenomic methods (more specifically ‘shotgun metagenomics’), where all the DNA in the sample is sheared, then the pieces are sequenced (Figure 2c) and analysed using computer software\textsuperscript{1,35} (Figure 2e). Metagenomic methods are capable of analyzing the entire collection of genomes and genes from all microbes present in the microbiome without relying on culture or amplification. The important advantage of metagenomic methods is the ability to quantitatively characterize the microbial communities in a sample by determining the relative abundance of and revealing potential interactions amongst species. Additionally, metagenomics allows for the characterization of genomic features (e.g., function) within the sample and allows for comparison across multiple samples and sample types. Metagenomic methods can easily detect the presence of viruses, fungus, and yet-to-be cultured bacteria. The use of complex computational algorithms needed in the analysis of metagenomic methods allow for a complete genomic analysis of the entire microbiome.

Although metagenomic methods show great promise in advancing the fields of biology and biomedical research, they are not without limitations. Despite the ability of the metagenomic
methods to infer the relative abundances of microbes (Figure 2e), such methods cannot
differentiate between living and dead microbes\textsuperscript{36} (although this can be achieved through
metatranscriptomics, but for typically double the cost) and have limited ability to detect microbes
that have low DNA representation in a sample. Additionally, the ability to study the microbiome
from human-dominated samples (e.g., tissue biopsy) is curbed as host DNA tends to overshadow
microbial reads\textsuperscript{3}. Further, clinical applications are complicated by the different metagenomic
prerequisites (Table 1; Figure 4a). Many challenges, however, lie with the computational aspect
of the methods, which have further variety even in run times processing the same data (Figure
5c).
Figure 5. Taxonomic Profiler Statistics. (a) A side by side comparison of sensitivity and precision at the genus level. There is no profiling technology with both precision and sensitivity both above 0.50. (b) A side by side comparison of sensitivity and precision at the species level. There is no mapping technology with precision and sensitivity both above 0.50. (c) CPU time, in hours, for each profiler to process the CAMI II Mouse Gut sample. CPU time data used from OPAL\textsuperscript{37}. Precision and Sensitivity data used from CAMI\textsuperscript{38}. 
The computational challenges and limitations of metagenomic methods

NGS technologies are capable of generating today’s “big data” sets across large-scale clinical cohorts. Large data sets generated by metagenomics methods require the use of sophisticated bioinformatics algorithms that are capable of differentiating technical noise from biological signals in the data and accurately assessing the metagenomic content of a given sample. However, several challenges must be addressed in order to leverage the full potential of metagenomic computational techniques in the clinical setting, including: lack of standardization in bioinformatics techniques, widely varying performance of such algorithms, and the lack of comprehensive reference databases (Figure 1a).

Developing an effective clinical diagnostic technique capable of properly informing medical decision-making requires specification of a standardized, detailed, and replicable procedures. In contrast, metagenomic computational techniques and pipelines are frequently published and distributed without standardized workflows, parameter settings, or input/output formats. As a result, metagenomic bioinformatics pipelines are frequently updated and evolve as new algorithms and tools are published. The ever increasing rate of newly developed computational methods and sequencing technologies causes the few existing published standardized procedures for metagenomic analysis (such as the admirable efforts of the Human Microbiome Project’s “Manual of Procedures”) to quickly become irrelevant. While some marker gene approaches, such as those using 16S rRNA, have begun to be standardized, there is less standardization and
consensus for whole genome shotgun (WGS) methods\textsuperscript{47}. Only recently have large-scale, unbiased assessments of individual WGS metagenomic tools been performed\textsuperscript{38,43,48,49}, with little in the way of standardizing entire computational pipelines. The academic software development community has made some effort to facilitate the ease of incorporating and assessing the performance of new methods in computational pipelines\textsuperscript{37,50–52}, but these efforts have not yet been widely adopted. This general lack of standardization results in many bioinformatics pipelines being created to analyze specific kinds of data resulting from measurements of specific biological samples using specific sequencing technology. Such a “bespoke” approach to software development limits a potentially wide adoption in the clinical setting where these exact conditions may not exist.

Today’s field of metagenomics research is dominated by a wide range of bioinformatics tools that have been published with very different performance characteristics. As a rapidly developing field, metagenomic bioinformatic tools, especially those in WGS metagenomics, still have room for improvement\textsuperscript{38,43,48,49}. One example can be found in metagenomic assembly — the process of building longer sequences from the shorter reads that are output from a sequencing machine, which can then be used for further analysis such as assessing the gene content of a metagenome. One study shows that current assemblers struggle to resolve individual strains from a metagenomic sample\textsuperscript{38,43}, a task important for detecting pathogens. Taxonomic profiling, detecting the presence and relative abundance of microbial taxa in a given sample, is another common computational technique for which there is little guidance on how to select the best tool. The consensus among all large-scale benchmarking studies is that selecting a single “best tool” is
often not possible or straightforward\textsuperscript{38,43,48,49}. This is due to some tools excelling at different specific metrics, including: sensitivity (correctly detecting microbial taxa that are actually present in a sample), specificity (correctly reporting as absent taxa that are not present in a sample), and correct prediction of the relative abundance of taxa. No current analytic technique excels at all three (or even two) of these metrics\textsuperscript{38,43} (Figure 5a; Figure 5b). Expert-level proficiency is often required in order to select the best possible tool for a given clinical application, and many medical institutions now tasked with metagenomics research lack access to computational expertise. In addition, metagenomic computational tools struggle with poor performance when asked to analyze samples at a resolution finer than the taxonomic level of genus\textsuperscript{38,43}. This is an unfortunate limitation when information at the level of species or strain is required in clinical settings.

Broad application of metagenomic computational techniques are presently limited in the clinical setting due to many incomplete, biased reference databases. Lack of comprehensive reference databases\textsuperscript{3,4,44,53} are another example of a need for a reference database of information concerning microbial organisms. Most metagenomic tools use a reference database comprised of information concerning microbial organisms in order to identify material from samples — including whole genome sequence databases\textsuperscript{54}, databases of gene families\textsuperscript{55}, and databases of taxonomic relationships\textsuperscript{56,57} (we refer to such information here collectively as “reference databases”). Currently available reference databases can be extremely variable in composition, and no single database represents the totality of existing information. For example, a recent analysis of fungal reference genome databases shows a greater than 30% discrepancy\textsuperscript{58} at the species level between
different reference databases. Similar discrepancies have been found in protein orthology
databases\textsuperscript{59} and in taxonomic databases\textsuperscript{60,61}. Furthermore, reference databases continue to
experience unprecedented growth in size\textsuperscript{62,63}. Some computational tools fail to utilize up-to-date
reference databases due to the time-consuming\textsuperscript{64} or difficult to implement nature of
post-processing procedures\textsuperscript{65}. The use of biased, incomplete, or outdated reference databases can
further negatively influence metagenomic computational algorithm performance. Expert-level
proficiency is often required to select for an application the reference database that is most
appropriate, up-to-date, and correct. The computational skills required to apply and interpret
metagenomics methods are not included in most graduate-level biological or biomedical
curricula\textsuperscript{66–68}.

**Leveraging metagenomics approaches to develop novel diagnostic tools for human disease**

There currently exists great potential for using metagenomic methods to develop tools capable of
detecting microbiota associated with human disease, infectious disease, and foodborne illness.
The microbiome has been proven to be an effective biomarker for conditions such as colorectal
cancer\textsuperscript{69–71}, inflammatory bowel disease (IBD) \textsuperscript{72,73}, and various metabolic syndromes\textsuperscript{74–76}.
Moreover, metagenomics can potentially be leveraged to produce applications that can be used to
determine the risk of infection during treatment procedures and can be used to develop
pre-operative procedures to reduce risk of complications. In some cases, microbiome analysis
may provide better insights into the etiology of disease then established diagnostic methods (e.g.,
blood sample analysis).
Infections within different tissue types pose unique challenges to the successful development of metagenomics-based methods to clinical settings. Microbiomes associated with the human body are sensitive to changes in diet\textsuperscript{77–81}, medications\textsuperscript{82–84}, nutritional supplements\textsuperscript{85}, environment\textsuperscript{86}, and health conditions\textsuperscript{87}. Microbiome compositions in the gut, skin, and other tissue types and organ systems can therefore be used as a predictor of illness. Colorectal cancer (CRC) -- a common cause of mortality across the globe -- has been linked to a noticeably altered microbial community within the intestines. Metagenomic analysis of gut microbiome in patients with CRC showed a reproducible increase in microbial species richness. Metagenomic analyses exploring microbial function and genomic content have established links between gluconeogenesis, putrefaction, and fermentation pathways in microbiota with CRC\textsuperscript{88}. Longitudinal studies have improved detection accuracy based on insights from combining global data from eight distinct geographical areas, which found that 29 core species were significantly enriched in the intestines of patients with CRC\textsuperscript{89}. CRC has also been shown to create an altered viral community within the intestines, which could also be detected using metagenomics methods\textsuperscript{90}. Furthermore, a number of tumor-associated microbiota have been linked to specific oral microbiota, suggesting that oral samples may provide an even easier sample site when compared to the gastrointestinal tract\textsuperscript{91}.

Analysis of the microbiome also shows strong potential for the diagnosis of inflammatory bowel disease (IBD), with the capacity to predict patient response to treatment and the chances of post-treatment relapse. Analysis of the gut microbiome in patients with IBD and irritable bowel
syndrome (IBS) show distinct microbial composition in the intestines of individuals who were diagnosed with the two conditions\textsuperscript{73}. Current diagnostic techniques for IBD and IBS often fail to distinguish between the two conditions, since they present similar symptomology in individuals suffering from chronic gastrointestinal distress. These studies have also led to the identification of key bacterial species with possible links to both IBD and IBS conditions\textsuperscript{73}.

Metagenomic-based approaches may allow physicians to distinguish between IBD and IBS and may provide researchers insight into the different causes of these diseases -- potentially leading to novel treatments for each. Beyond the diagnostic potential of the microbiota in IBD, metagenomics has also provided clinicians and researchers with a roadmap for future intervention studies. Using metagenomics to link IBD severity with microbiota structure and diet, several studies have already begun tailoring dietary and nutritional interventions for patients with IBD\textsuperscript{92,93}.

The strategy of linking metagenomics with other metadata and variables, like diet, has strong potential beyond gastrointestinal disorders. In addition to diagnosis of a known disease, metagenomics provides a unique opportunity to explore diseases of unknown etiology. This can be particularly helpful in the face of emerging, zoonotic, or rare infectious agents. Furthermore, metagenomics has provided the opportunity to discover previously unknown and unstudied classes of viruses, which may have unappreciated roles in disease\textsuperscript{94}. Additionally, metagenomics allows the opportunity to study the interactions of different microorganisms (bacteria, virus, eukaryotic microbes) to produce disease phenotypes. This is, in fact, a paradigm shift from the last century of detecting/diagnosing ‘the’ organism causing disease. Metagenomic techniques
allow researchers to characterize the entire microbiome and the relative contributions of 
individual components and interaction effects towards disease phenotypes.

This application to phenotypes is not limited to disease, but also phenotypes associated with 
mental health issues. For example, human genome analysis is often used when studying the 
development of schizophrenia\textsuperscript{95}, however, in a recent 2018 study utilizing blood samples, 
patients with schizophrenia are shown to have higher microbiome diversity than others with 
amyotrophic lateral sclerosis and bipolar disorder, and healthy controls\textsuperscript{96}. The gut microbiome of 
schizophrenia patients, too, differed from gut microbiomes of healthy control groups\textsuperscript{97}. There is 
evidence that the gut microbiome is capable of affecting behavior associated with 
schizophrenia\textsuperscript{97} with additional studies suggesting this could occur through the 
microbiota-gut-brain axis (MGB)\textsuperscript{98}, though relationships between the MGB and schizophrenia 
are not well studied. The connection between host-associated microbiota and mental health 
extend beyond schizophrenia to phenotypes for anxiety and trauma-related disorders\textsuperscript{99}.

Metagenomics-based studies also have potential to revolutionize the way well-understood 
bacterial infections are identified, affecting treatment plans. When treating bacterial infections, 
clinics typically prescribe broad spectrum antibiotics. Clinicians could send a sample of infected 
tissue to a lab for a more specific diagnosis; traditionally, a lab uses culture-based methods to 
identify the infecting agent in order to choose a more specific antibiotic to prescribe. However, 
this process can take multiple days, during which time the broad-spectrum antibiotics can harm 
the natural microbiome of the patient, worsen symptoms, or have no effect on the infection,
allowing it to continue spreading in the patient’s body. A culture-independent technique using diagnostic PCR, which is used to amplify the DNA of specific microbes known to cause certain diseases, can fail to identify the harmful biota. Metagenomic methods can provide results in hours (not days), and with the use of amplification techniques of untargeted sequences, a patient can get treatment specific to the composition of the infection. These amplification techniques do incur biases in the resulting sequences, and caution should be used when utilizing them\(^{100}\).

Ultimately, metagenomics can help a patient avoid the harmful effects of broad-spectrum antibiotic treatment and potentially limit the evolution of new antibiotic resistant bacterial strains.

Metagenomics can provide more information than simply identification of the cause of diseases and infections -- it also has potential for early diagnostic prior to an individual’s manifestation of noticeable symptoms. Research in this area is especially important for cancer patients because chemotherapy treatments weaken the immune response to foreign pathogens. A massive longitudinal study found that a decrease in stool microbial richness is associated with infection while receiving induction chemotherapy treatments for leukemia\(^{101}\). An additional study found that cancer patients receiving Hematopoietic stem cell transplantation are at an increased risk for bloodstream infection if the gut microbiome, before chemotherapy, shows decreased diversity and lowered abundance of certain taxa\(^{102}\).

Metagenomics studies also show promise in the development of strain identification for *Mycobacterium tuberculosis* (TB) -- which has an emergent strain showing drug resistance to
isoniazid and rifampicin, called MDR-TB. Today’s available treatment plans for additional strains, resistant to fluoroquinolones and other antibiotics, called XDR-TB, have only a 19% favorable outcome\textsuperscript{103}. Current analysis of globally reported cases of TB show 4% of cases being rifampicin resistant\textsuperscript{104}. Drug resistance for cases of TB were traditionally determined using CDMs, whereby the infecting strain is grown in the presence of various antibiotics. A CIM, utilizing GeneXpert technology, has been implemented, and it can detect cases of TB on site and identify rifampicin resistance within two hours\textsuperscript{105}. The results of this test can determine whether or not a patient should receive treatment for MDR-TB or the standard TB treatment. Although reliable for fully evolved MDR-TB strains, recent genetic analysis of the evolution of TB drug resistance have shown that the development of resistance to isoniazid occurs before the development of resistance to rifampicin\textsuperscript{103}. Patients infected with TB at earlier stages of drug resistant evolution, resistant to isoniazid but not rifampicin, would not be identified as MDR-TB by the current CIMs. The treatment of these cases with the standard TB treatment allows for the further evolution of antibiotic resistance into MDR-TB, thus increasing its prevalence.

*Mycobacterium tuberculosis* is not the only bacterium in need for more precise monitoring techniques. Harmful bacteria found in farm animals and meat produced for human consumption have developed a resistance to Tigecycline (an antibiotic of last resort to treat severe infections). The current cause of this resistance is associated with the evolution of two genes named *tet*(X3) and *tet*(X4), which inactivate all tetracyclines (the classification of antibiotics under which Tigecycline falls). Even newly FDA-approved antibiotics, including eravacycline and omadacycline, are also ineffective. These two genes are found on plasmids and can therefore be
copied and given to other bacteria through horizontal gene transfer. The spread of resistant bacteria and their genes have not been fully quantified; however, resistant strains have already been found in hospital patients in China\textsuperscript{106}. Metagenomic-based methods have the ability to quickly determine the resistance of a bacterial strain before the administration of potentially ineffective drugs -- a development key to saving the lives of thousands of individuals affected by antibiotic-resistant strains of harmful bacteria.

Encephalitis is an inflammation of the brain caused by an infection that, if left untreated, can result in serious disability and carries a 30% mortality rate. Currently, available reference databases for encephalitis are incomplete; identification of the bacterial agent is typically performed using diagnosis PCR, which only amplifies the DNA of microbes already known to cause encephalitis. Currently, the instigating microbe in more than 33% of encephalitis cases are unidentified. NGS based methods are capable of identifying the cause of encephalitis, and in the process, identify strains and organisms not previously associated with the condition. Therefore, metagenomics methods are capable of identifying malicious microbes (even those never associated with harmful infections)\textsuperscript{107}.

However, there are unique challenges associated with detecting infectious diseases from different human tissue types and a potential necessity for enrichment protocols designed to remove host DNA. Sources for studying the microbiome within humans come in a variety of forms, each of which have their own difficulties during DNA processing prior to sequencing. These sources include solids, such as feces; tissues, such as colon, cancer cells, or other organ samples; swabs,
such as from sampling the skin, oral, nasal or vaginal epithelial cells; and liquids, such as saliva, breast milk, urine, or blood. Samples taken from tissues and swabs tend to be oversaturated with the individual’s own human DNA (>90% human genome-aligned reads for swab samples\textsuperscript{108}). This oversaturation can obscure the infecting microbe(s) and increase the likelihood of false negatives. There are two general ways to remove host DNA from a sample - either depletion of host cells or enrichment of microbial cells\textsuperscript{109,110}. Although these techniques sound similar, the approaches are different. Depletion of human DNA is completed pre-extraction (\textbf{Figure 2b}) by a selective lysis and degradation of the eukaryotic (e.g., human but also fungal and protist) cells followed by purification of the sample to remove the degraded cells. This process leaves the intact bacterial cells behind to be extracted. Recent work has been able to remove as much as 99% of human DNA from lung tissue before the PCR process\textsuperscript{111,112}. Alternatively, enrichment of microbial cells is completed post-extraction (\textbf{Figure 2b}) by a methylation selection, where CpG-methylated host DNA is selectively bound and the microbial DNA is eluted\textsuperscript{110}. For example, there was a 50-fold decrease in reads aligning to host genomes when using this method\textsuperscript{110}. Both methods have a potential to remove microbes from the sample during this step; host depletion will lyse fungal and protist cells and may also lyse bacterial cells, and not all microbial cells are methylation free. Therefore, using host depletion techniques prior to metagenomic methods may prevent the detection of parasitic protists, such as those that cause malaria or giardiasis. Fluids, as well, tend to be oversaturated with human DNA (>90% saliva metagenomic reads mapped to human genome\textsuperscript{108}) but are also severely diluted due to their nature, which results in extremely low yields of microbial DNA within the sample. To circumvent this, fluids have been mostly studied by metataxonomic methods (i.e., 16S rRNA
amplicon sequencing). However, efforts are being made to adapt metagenomic methods and microbial enrichment techniques to study these samples. For example, a microbial enrichment method increased the number of mapped reads to the *Neisseria gonorrhoeae* genome from 1% up to 43% in clinical samples of urine that tested positive for *N. gonorrhoeae*\(^{13}\). The most common source of starting material for human metagenomic studies is fecal samples, which is often used as a proxy for studying the gut microbiome. Unlike other sample sources, fecal samples have less host DNA contamination\(^ {114}\) (<10% metagenomic reads mapped to the human genome\(^ {108}\) ) and depletion of host DNA is not needed. In a recent clinical study which used fecal samples to study the longitudinal effect of a fecal transplant in children with *Clostridioides difficile* did not complete any host depletion and 98.26% of all reads mapped to bacterial genomes with 1.66% of all reads mapped to viruses and archaea\(^ {115}\).

There are other diagnostic applications outside of those mentioned. A review of the current clinical metagenomic studies in the last five years showed sample sizes varying from 1 to 204 with both fresh and frozen samples and sequencing depth as low as .0002 and as high as 9.75 Gbp per sample. The main sequencing technology being used is Illumina. A variety of bioinformatics tools are in use with the most popular being MetaPhlAn2 used in 55% of selected clinical metagenomic studies (Table 2).

**Table 2: An overview of clinical metagenomics studies published between 2015-2019.** Ref is the citation for the clinical metagenomics paper. Host-DNA depletion is yes/no do they process the sample to remove host DNA before sequencing. Sequencing technology is the sequencing
technology used to sequence the sample. Tissue is the source of the sample from the subject.

**Sequencing Depth** is given in Gbps. This is the throughput of each study, an ‘X’ is given when the data is not publicly available for analysis. **Disease** is the ailment that the study focuses on.

**Data public** are the raw reads from the study publically available (yes/no). **Sample size (n)** is the number of total participants in the study. **Bioinformatics tools** are the computer programs used to analyse the sequenced data. **Fresh or frozen samples** tells whether the samples taken were frozen before sequencing or used fresh.

*Sputum refers to saliva and mucus that is eliminated from the respiratory tract.

| Ref | Host-DNA depletion | Sequencing technology | Sequencing depth (Gbp) | Tissue | Disease | Data public | Sample size (n) | Bioinformatics tools | Fresh or frozen samples |
|-----|-------------------|-----------------------|------------------------|--------|---------|-------------|------------------|----------------------|------------------------|
| 116 | Yes               | Illumina HiSeq 2500   | X                      | Bone and joint infections | No      | 47         |                  | MetaPhlAn2 Kraken Bowtie2 | Frozen                |
| 117 | Yes               | Illumina MiSeq        | 2                      | Aortic valve              | Infective endocarditis | Yes | 1          | CLARK MetaPhlAn2 mothur UBLAST | Frozen                |
| 118 | No                | Illumina MiSeq        | 0.041                  | Aortic valve              | Infective endocarditis | Yes | 1          | CLARK USEARCH MetaPhlAn2 BWA BLAST | Frozen                |
| 119 | No                | Illumina MiSeq        | N/A                    | Cerebrospinal fluid       | Toscana virus          | Yes¹ | 1          | TOSV SmaltAlign VirMet | Fresh                 |
| 120 | No                | Illumina GAIIx        | 0.07                   | Sputum*                   | Polymicrobial infections | Yes | 1          | SMALT BLASTn | Frozen                |
| # | Is | Runners | Read length | Sample type | Organism | Genomic data | Bioinformatic tools | Preparation state |
|---|---|---|---|---|---|---|---|---|
| 121 | Yes | Illumina NextSeq | 9.75 | Feces and rectal swabs | Klebsiella pneumoniae | Yes | 2 | Fresh |
| 122 | Yes | Illumina MiSeq | 0.0002 | Liver | Chronic hepatic brucellosa | Yes | 1 | Frozen |
| 123 | Yes | Illumina HiSeq 2000 and 2500 | 1 | Broncho-alveolar lavage | Chronic pneumonia | Yes | 1 | Frozen |
| 124 | Yes | Illumina HiSeq X | 1 | Cerebrospinal Fluid | Meningitis and encephalitis | No | 204 | Frozen |
| 125 | Yes | Nanopore MinION | N/A | Sputum, bronchoalveolar lavage (BAL) and endotracheal aspirates (ETAs) | Lower respiratory infections | Yes | 40 | Frozen |
| 126 | No | Illumina HiSeq 2500 | 1.85 | Resected prostheses | Prosthetic joint infections | Yes | N/A | Fresh |

Leveraging the potential of metagenomics approaches for therapeutic applications of microbial research
The microbiome has a strong effect on an individual’s ability to interact with, excrete, or metabolize food and medication. Understanding the impact that the microbiome has on the activity of pharmaceutical drugs will become essential for clinicians to develop personalized therapies for patients. Understanding the link between an individual’s microbiome and metabolism, weight gain, and metabolic diseases can also be used in personalized medicine. Specifically, metagenomics-based approaches show promise in predicting an individual’s probability of developing melanoma and IBD, as well as determining the amount of a medication that will actually be processed in a patient’s body (bioavailability). Metagenomics has a clear role in potential development of microbiome-based treatment -- including treating individuals with unhealthy gut microbiota, obesity, and diabetes. Protocols for this precision metagenomics have already been proposed\textsuperscript{127}.

Microbes (such as those in the phylum Bacteroidetes) have even been shown to interfere with the liver’s decomposition of widely used drugs, such as acetaminophen, leading to the increased formation of more toxic and carcinogenic compounds\textsuperscript{6}. Additionally, a decrease in Bacteroidetes abundance has been seen in women who have indicators of blood glucose control issues\textsuperscript{128–130}. Aside from modulating drug activity, the most important potential applications can be found in the administration of cancer treatments. The presence of certain microbes can increase or decrease the toxicity or efficacy of treatment\textsuperscript{7}. Having an outline of the microbes present inside a patient can inform doctors on the appropriate course of drugs.
There are additional applications of microbiome-based treatment (including treating individuals having unhealthy gut microbiota) and the role of metagenomics in developing such applications is a topic of current study. These studies have led to the trial use of microbiome transplants as a treatment plan for various infections and conditions and have the potential to replace standard antibacterial treatments that are quickly becoming ineffective\textsuperscript{44,131–133}. Fecal Microbiota Transplants (FMT) have emerged as a viable treatment for \textit{Clostridioides difficile} infection (CDI). As CDI increases in prevalence\textsuperscript{134–137}, the risk for recurrent infection is also escalating, with a 20-60% chance of experiencing a CDI again\textsuperscript{138,139}. CDI drastically decreases the fecal bacterial diversity in an affected person, and FMT is considered an effective treatment in adults, although the adverse effects are not well known or characterized. Recently, the death of a patient caused the FDA to issue a safety alert regarding the transfer of multidrug-resistant organisms by FMT\textsuperscript{140}, which reinforces the importance of screening the donor prior to administering a FMT to the patient. Metagenomic methods facilitated an in depth investigation of the longitudinal changes on children’s gut microbiome pre- and post-FMT and the potential acquisition of antimicrobial resistant genes (AMR) and pathogens from their adult donors\textsuperscript{141}. In this study, FMT reduced AMR genes and potential pathogens within the children, and the microbial diversity increased post-FMT and was sustained over time\textsuperscript{141}. Because of the resolution that shotgun metagenomics provides, bacterial species that are in high abundance in the human gut\textsuperscript{142} were shown to increase significantly post-FMT; this difference, along with other changes in abundances of species, was not detected when using 16S rRNA sequencing\textsuperscript{143}. The efficacy of this treatment strategy for more complex microbiota-associated gastrointestinal disorders remains to be tested but trials are currently ongoing to test the application of FMT in IBD, IBS,
and obesity\cite{144}. Additionally, FMT has also been considered as a treatment for neuropsychiatric disorders, such as autism in which patients treated with FMT experienced improved symptoms\cite{144,145}. As with CDI, the functional capability that was changed post-FMT\cite{141} can also be investigated with these other, more complex disorders with the use of metagenomics.

Connections between the human microbiome and various metabolic functions have been the topic of study for a number of years. In a recent study using \textit{Akkermansia muciniphila} as a supplement for three months, patients saw an increase in insulin sensitivity, and a slight decrease in body weight ($-2.27 \pm 0.92$ kg, $P = 0.091$) compared to the control (which took a placebo)\cite{146}.

\textbf{Discussion}

Metagenomic methods have a wide range of utility and applications. These include, but are not limited to, infection detection, surveillance and tracking of outbreaks, antimicrobial resistance prediction and strain resolution, and disease prognosis.

Overcoming the current limitations of metagenomics methods is necessary not only for the field of microbiology to progress, but also to overcome the major hurdle of acquiring clinical validation. The method must be reproducible, reliable, and accurate to be clinically applicable\cite{147}, which most metagenomic tools are not. As of writing this paper, there are only three validated clinical tools based on metagenomics methods. The first is the use of cerebrospinal fluid to identify the infectious agent in encephalitis and meningitis cases. This is done by the Clinical
Microbiology Lab at UC San Francisco. The second is the use of bronchoalveolar lavage fluid for lower respiratory tract infections. This is done by IDbyDNA. Finally, identification of microbial cell-free DNA from blood plasma is done by Karius Inc. All of these methods underwent extreme scrutiny and have published papers providing evidence that their method is reproducible, reliable, and accurate\textsuperscript{148--150}.

Clinical validation also depends on financial interest in the medical research and development community. Insurance companies traditionally have not covered the costs of metagenomic tests for individuals who wish to use available novel methods. This is because there is not yet enough data showing that these methods can effectively improve patient outcomes. There need to be larger studies and pilot programs that can demonstrate that not only are these methods effective, but that are capable of saving lives and money while also reducing the need to depend on general antibiotics both as a health issue and as antimicrobial resistance continues.
References Cited

1. Marchesi, J. R. & Ravel, J. The vocabulary of microbiome research: a proposal.
   *Microbiome* 3, 31 (2015).

2. Lagier, J.-C. *et al.* Current and past strategies for bacterial culture in clinical microbiology.
   *Clin. Microbiol. Rev.* 28, 208–236 (2015).

3. Forbes, J. D., Knox, N. C., Ronholm, J., Pagotto, F. & Reimer, A. Metagenomics: The Next
   Culture-Independent Game Changer. *Front. Microbiol.* 8, 1069 (2017).

4. Kyrpides, N. C., Eloe-Fadrosh, E. A. & Ivanova, N. N. Microbiome Data Science:
   Understanding Our Microbial Planet. *Trends Microbiol.* 24, 425–427 (2016).

5. Human Microbiome Project Consortium. Structure, function and diversity of the healthy
   human microbiome. *Nature* 486, 207–214 (2012).

6. Spanogiannopoulos, P., Bess, E. N., Carmody, R. N. & Turnbaugh, P. J. The microbial
   pharmacists within us: a metagenomic view of xenobiotic metabolism. *Nature Reviews
   Microbiology* 14, 273–287 (2016).

7. Helmink, B. A., Wadud Khan, M. A., Hermann, A., Gopalakrishnan, V. & Wargo, J. A. The microbiome, cancer, and cancer therapy. *Nature Medicine* 25, 377–388 (2019).

8. Escobar-Zepeda, A., Vera-Ponce de León, A. & Sanchez-Flores, A. The Road to
   Metagenomics: From Microbiology to DNA Sequencing Technologies and Bioinformatics.
   *Front. Genet.* 6, 348 (2015).

9. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating
   inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* 74, 5463–5467 (1977).

10. Woese, C. R. & Fox, G. E. Phylogenetic structure of the prokaryotic domain: The primary
kingdoms. *Proceedings of the National Academy of Sciences* **74**, 5088–5090 (1977).

11. Shine, J. & Dalgarno, L. The 3’-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc. Natl. Acad. Sci. U. S. A.* **71**, 1342–1346 (1974).

12. Dahlberg, A. E. The functional role of ribosomal RNA in protein synthesis. *Cell* **57**, 525–529 (1989).

13. Hagenbüchle, O., Santer, M., Steitz, J. A. & Mans, R. J. Conservation of the primary structure at the 3’ end of 18S rRNA from eucaryotic cells. *Cell* **13**, 551–563 (1978).

14. Fox, G. E., Woese, C. R. & Pechman, K. R. Comparative Cataloging of 16S Ribosomal Ribonucleic Acid: Molecular Approach to Procaryotic Systematics. *International Journal of Systematic and Evolutionary Microbiology* **27**, 44–57 (1977).

15. Van de Peer, Y., Chapelle, S. & De Wachter, R. A quantitative map of nucleotide substitution rates in bacterial rRNA. *Nucleic Acids Res.* **24**, 3381–3391 (1996).

16. Ferri, G., Alù, M., Corradini, B., Licata, M. & Beduschi, G. Species Identification Through DNA ‘Barcodes’. *Genetic Testing and Molecular Biomarkers* **13**, 421–426 (2009).

17. Chakravorty, S., Helb, D., Burday, M., Connell, N. & Alland, D. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *Journal of Microbiological Methods* **69**, 330–339 (2007).

18. Pereira, F. *et al.* Identification of species by multiplex analysis of variable-length sequences. *Nucleic Acids Res.* **38**, e203 (2010).

19. DeSantis, T. Z. *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72**, 5069–5072 (2006).
20. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–6 (2013).

21. lexnederbragt. Developments in high throughput sequencing – July 2016 edition. *In between lines of code* (2016). Available at: https://flxlexblog.wordpress.com/2016/07/08/developments-in-high-throughput-sequencing-july-2016-edition/. (Accessed: 15th October 2019)

22. Leland, D. S. & Ginocchio, C. C. Role of cell culture for virus detection in the age of technology. *Clin. Microbiol. Rev.* **20**, 49–78 (2007).

23. Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107**, 1–8 (2016).

24. Saiki, R. K. et al. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239**, 487–491 (1988).

25. Peake, I. The polymerase chain reaction. *J. Clin. Pathol.* **42**, 673–676 (1989).

26. Wong, C. et al. Characterization of beta-thalassaemia mutations using direct genomic sequencing of amplified single copy DNA. *Nature* **330**, 384–386 (1987).

27. Ieven, M. & Ursi, D. [Application of polymerase chain reaction in medical microbiology]. *Verh. K. Acad. Geneeskd. Belg.* **56**, 593–623 (1994).

28. Hartskeerl, R. A., De Wit, M. Y. L. & Klatser, P. R. Polymerase Chain Reaction for the Detection of Mycobacterium leprae. *Microbiology* **135**, 2357–2364 (1989).

29. de Wit, M. Y. et al. Application of a polymerase chain reaction for the detection of Mycobacterium leprae in skin tissues. *J. Clin. Microbiol.* **29**, 906–910 (1991).

30. Singh, P. & Cole, S. T. Mycobacterium leprae: genes, pseudogenes and genetic diversity.
31. Flaman, J.-M. et al. A rapid PCR fidelity assay. *Nucleic Acids Research* **22**, 3259–3260 (1994).

32. Mullis, K. et al. Specific Enzymatic Amplification of DNA In Vitro: The Polymerase Chain Reaction. *Cold Spring Harbor Symposia on Quantitative Biology* **51**, 263–273 (1986).

33. Bybee, S. M. et al. Targeted amplicon sequencing (TAS): a scalable next-gen approach to multilocus, multitaxa phylogenetics. *Genome Biol. Evol.* **3**, 1312–1323 (2011).

34. Hodgkin, J. *Encyclopedia of Genetics*. 865 (Academic Press, 2001).

35. Sharpton, T. J. An introduction to the analysis of shotgun metagenomic data. *Frontiers in Plant Science* **5**, (2014).

36. Zapka, C. et al. Comparison of Standard Culture-Based Method to Culture-Independent Method for Evaluation of Hygiene Effects on the Hand Microbiome. *MBio* **8**, (2017).

37. Meyer, F. et al. Assessing taxonomic metagenome profilers with OPAL. *Genome Biol.* **20**, 51 (2019).

38. Sczyrba, A. et al. Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software. *Nat. Methods* **14**, 1063–1071 (2017).

39. Iñiguez, L. P. et al. Transcriptomic analysis of human endogenous retroviruses in systemic lupus erythematosus. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 21350–21351 (2019).

40. Zeevi, D. et al. Structural variation in the gut microbiome associates with host health. *Nature* **568**, 43–48 (2019).

41. Feinstein, A. R. Clinical epidemiology: The architecture of clinical research. *Stat. Med.* **14**, 1263–1263 (1995).
42. Sackett, D. L., Haynes, R. B., Tugwell, P. & Guyatt, G. H. *Clinical epidemiology: a basic science for clinical medicine*. (Little, Brown Boston, 1985).

43. Ye, S. H., Siddle, K. J., Park, D. J. & Sabeti, P. C. Benchmarking Metagenomics Tools for Taxonomic Classification. *Cell* **178**, 779–794 (2019).

44. Chiu, C. Y. & Miller, S. A. Clinical metagenomics. *Nat. Rev. Genet.* (2019).
   doi:10.1038/s41576-019-0113-7

45. Aagaard, K. *et al.* The Human Microbiome Project strategy for comprehensive sampling of the human microbiome and why it matters. *FASEB J.* **27**, 1012–1022 (2013).

46. D’Amore, R. *et al.* A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. *BMC Genomics* **17**, 55 (2016).

47. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).

48. Lindgreen, S., Adair, K. L. & Gardner, P. P. An evaluation of the accuracy and speed of metagenome analysis tools. *Sci. Rep.* **6**, 19233 (2016).

49. McIntyre, A. B. R. *et al.* Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome Biol.* **18**, 182 (2017).

50. Meyer, F. *et al.* AMBER: Assessment of Metagenome BinnERs. *Gigascience* **7**, (2018).

51. Belmann, P. *et al.* Bioboxes: standardised containers for interchangeable bioinformatics software. *Gigascience* **4**, 47 (2015).

52. Mason, C. E., Afshinnekoo, E., Tighe, S., Wu, S. & Levy, S. International Standards for Genomes, Transcriptomes, and Metagenomes. *J. Biomol. Tech.* **28**, 8–18 (2017).

53. Almeida, A. *et al.* A new genomic blueprint of the human gut microbiota. *Nature* (2019).
doi:10.1038/s41586-019-0965-1

54. Tatusova, T., Ciufo, S., Fedorov, B., O’Neill, K. & Tolstoy, I. RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res.* **42**, D553–9 (2014).

55. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

56. Yilmaz, P. *et al.* The SILVA and ‘All-species Living Tree Project (LTP)’ taxonomic frameworks. *Nucleic Acids Res.* **42**, D643–D648 (2014).

57. Federhen, S. The NCBI Taxonomy database. *Nucleic Acids Res.* **40**, D136–43 (2012).

58. Loeffler, C., Karlsberg, A., Eskin, E., Koslicki, D. & Mangul, S. Analysis of multiple fungal sequence repositories highlights shortcomings in microbial databases.
doi:10.1101/497867

59. Alexeyenko, A., Lindberg, J., Pérez-Bercoff, A. & Sonnhammer, E. L. L. Overview and comparison of ortholog databases. *Drug Discov. Today Technol.* **3**, 137–143 (2006).

60. Balvočiūtė, M. & Huson, D. H. SILVA, RDP, Greengenes, NCBI and OTT—how do these taxonomies compare? *BMC Genomics* **18**, 114 (2017).

61. Breitwieser, F. P., Lu, J. & Salzberg, S. L. A review of methods and databases for metagenomic classification and assembly. *Brief. Bioinform.* (2017). doi:10.1093/bib/bbx120

62. Cook, C. E. *et al.* The European Bioinformatics Institute in 2016: Data growth and integration. *Nucleic Acids Res.* **44**, D20–6 (2016).

63. Galperin, M. Y., Fernández-Suárez, X. M. & Rigden, D. J. The 24th annual Nucleic Acids Research database issue: a look back and upcoming changes. *Nucleic Acids Res.* **45**,
64. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).

65. Truong, D. T. *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).

66. Welch, L. *et al.* Bioinformatics Curriculum Guidelines: Toward a Definition of Core Competencies. *PLoS Computational Biology* **10**, e1003496 (2014).

67. Mangul, S., Martin, L. S., Hoffmann, A., Pellegrini, M. & Eskin, E. Addressing the Digital Divide in Contemporary Biology: Lessons from Teaching UNIX. *Trends Biotechnol.* **35**, 901–903 (2017).

68. Mangul, S. Interpreting and integrating big data in the life sciences. doi:10.7287/peerj.preprints.27603v1

69. Zackular, J. P., Rogers, M. A. M., Ruffin, M. T., 4th & Schloss, P. D. The human gut microbiome as a screening tool for colorectal cancer. *Cancer Prev. Res.* **7**, 1112–1121 (2014).

70. Baxter, N. T., Ruffin, M. T., 4th, Rogers, M. A. M. & Schloss, P. D. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome Med.* **8**, 37 (2016).

71. Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* **10**, 766 (2014).

72. Zhou, Y. *et al.* Gut Microbiota Offers Universal Biomarkers across Ethnicity in Inflammatory Bowel Disease Diagnosis and Infliximab Response Prediction. *mSystems* **3**,
73. Vich Vila, A. et al. Gut microbiota composition and functional changes in inflammatory bowel disease and irritable bowel syndrome. *Sci. Transl. Med.* 10, (2018).

74. Yang, T. et al. Gut Dysbiosis Is Linked to Hypertension. *Hypertension* **65**, 1331–1340 (2015).

75. Liou, A. P. et al. Conserved shifts in the gut microbiota due to gastric bypass reduce host weight and adiposity. *Sci. Transl. Med.* **5**, 178ra41 (2013).

76. Roopchand, D. E. et al. Dietary Polyphenols Promote Growth of the Gut Bacterium Akkermansia muciniphila and Attenuate High-Fat Diet-Induced Metabolic Syndrome. *Diabetes* **64**, 2847–2858 (2015).

77. David, L. A. et al. Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**, 559–563 (2014).

78. Wu, G. D. et al. Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334**, 105–108 (2011).

79. Duncan, S. H. et al. Reduced dietary intake of carbohydrates by obese subjects results in decreased concentrations of butyrate and butyrate-producing bacteria in feces. *Appl. Environ. Microbiol.* **73**, 1073–1078 (2007).

80. Walker, A. W. et al. Dominant and diet-responsive groups of bacteria within the human colonic microbiota. *ISME J.* **5**, 220–230 (2011).

81. King, C. H. et al. Baseline human gut microbiota profile in healthy people and standard reporting template. *PLoS One* **14**, e0206484 (2019).

82. Rogers, M. A. M. & Aronoff, D. M. The influence of non-steroidal anti-inflammatory drugs
on the gut microbiome. *Clin. Microbiol. Infect.* **22**, 178.e1–178.e9 (2016).

83. Forslund, K. *et al.* Corrigendum: Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature* **545**, 116 (2017).

84. Maier, L. *et al.* Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature* **555**, 623–628 (2018).

85. Zackular, J. P. *et al.* Dietary zinc alters the microbiota and decreases resistance to Clostridium difficile infection. *Nat. Med.* **22**, 1330–1334 (2016).

86. Spor, A., Koren, O. & Ley, R. Unravelling the effects of the environment and host genotype on the gut microbiome. *Nat. Rev. Microbiol.* **9**, 279–290 (2011).

87. Stiensma, L. T. *et al.* Shifts in Lachnospira and Clostridium sp. in the 3-month stool microbiome are associated with preschool age asthma. *Clin. Sci.* **130**, 2199–2207 (2016).

88. Thomas, A. M. *et al.* Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nature Medicine* **25**, 667–678 (2019).

89. Wirbel, J. *et al.* Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med.* **25**, 679–689 (2019).

90. Hannigan, G. D., Duhaime, M. B., Ruffin, M. T., 4th, Koumpouras, C. C. & Schloss, P. D. Diagnostic Potential and Interactive Dynamics of the Colorectal Cancer Virome. *MBio* **9**, (2018).

91. Flemer, B. *et al.* The oral microbiota in colorectal cancer is distinctive and predictive. *Gut* **67**, 1454–1463 (2018).

92. Lee, D. *et al.* Comparative Effectiveness of Nutritional and Biological Therapy in North
American Children with Active Crohn’s Disease. *Inflamm. Bowel Dis.* **21**, 1786–1793 (2015).

93. Sartor, R. B. & Wu, G. D. Roles for Intestinal Bacteria, Viruses, and Fungi in Pathogenesis of Inflammatory Bowel Diseases and Therapeutic Approaches. *Gastroenterology* **152**, 327–339.e4 (2017).

94. Abbas, A. A. *et al.* Redondoviridae, a Family of Small, Circular DNA Viruses of the Human Oro-Respiratory Tract Associated with Periodontitis and Critical Illness. *Cell Host Microbe* **25**, 719–729.e4 (2019).

95. Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium. Genome-wide association study identifies five new schizophrenia loci. *Nat. Genet.* **43**, 969–976 (2011).

96. Olde Loohuis, L. M. *et al.* Transcriptome analysis in whole blood reveals increased microbial diversity in schizophrenia. *Transl. Psychiatry* **8**, 96 (2018).

97. Correction for the Research Article: ‘The gut microbiome from patients with schizophrenia modulates the glutamate-glutamine-GABA cycle and schizophrenia-relevant behaviors in mice’ by P. Zheng, B. Zeng, M. Liu, J. Chen, J. Pan, Y. Han, Y. Liu, K. Cheng, C. Zhou, H. Wang, X. Zhou, S. Gui, S. W. Perry, M. Wong, J. Licinio, H. Wei, and P. Xie. *Sci Adv* **5**, eaay2759 (2019).

98. Cryan, J. F. & Dinan, T. G. Mind-altering microorganisms: the impact of the gut microbiota on brain and behaviour. *Nat. Rev. Neurosci.* **13**, 701–712 (2012).

99. Malan-Muller, S. *et al.* The Gut Microbiome and Mental Health: Implications for Anxiety- and Trauma-Related Disorders. *OMICS* **22**, 90–107 (2018).
100. Ahsanuddin, S. et al. Assessment of REPLI-g Multiple Displacement Whole Genome Amplification (WGA) Techniques for Metagenomic Applications. *J. Biomol. Tech.* **28**, 46–55 (2017).

101. Galloway-Peña, J. R. et al. The role of the gastrointestinal microbiome in infectious complications during induction chemotherapy for acute myeloid leukemia. *Cancer* **122**, 2186–2196 (2016).

102. Montassier, E. et al. Erratum to: Pretreatment gut microbiome predicts chemotherapy-related bloodstream infection. *Genome Med.* **8**, 61 (2016).

103. Manson, A. L. et al. Genomic analysis of globally diverse *Mycobacterium tuberculosis* strains provides insights into the emergence and spread of multidrug resistance. *Nat. Genet.* **49**, 395–402 (2017).

104. Feyisa, S. G. et al. Resistance of *Mycobacterium tuberculosis* strains to Rifampicin: A systematic review and meta-analysis. *Heliyon* **5**, e01081 (2019).

105. Saeed, M. et al. GeneXpert technology. A breakthrough for the diagnosis of tuberculous pericarditis and pleuritis in less than 2 hours. *Saudi Medical Journal* **38**, 699–705 (2017).

106. He, T. et al. Emergence of plasmid-mediated high-level tigecycline resistance genes in animals and humans. *Nat Microbiol* (2019). doi:10.1038/s41564-019-0445-2

107. Brown, J. R., Bharucha, T. & Breuer, J. Encephalitis diagnosis using metagenomics: application of next generation sequencing for undiagnosed cases. *J. Infect.* **76**, 225–240 (2018).

108. Marotz, C. A. et al. Improving saliva shotgun metagenomics by chemical host DNA depletion. *Microbiome* **6**, 42 (2018).
109. Bachmann, N. L., Rockett, R. J., Timms, V. J. & Sintchenko, V. Advances in Clinical Sample Preparation for Identification and Characterization of Bacterial Pathogens Using Metagenomics. *Front Public Health* **6**, 363 (2018).

110. Feehery, G. R. *et al.* A method for selectively enriching microbial DNA from contaminating vertebrate host DNA. *PLoS One* **8**, e76096 (2013).

111. Charalampous, T. *et al.* Rapid Diagnosis of Lower Respiratory Infection using Nanopore-based Clinical Metagenomics. doi:10.1101/387548

112. Charalampous, T. *et al.* Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. *Nature Biotechnology* (2019). doi:10.1038/s41587-019-0156-5

113. Graham, R. M. A., Doyle, C. J. & Jennison, A. V. Epidemiological typing of Neisseria gonorrhoeae and detection of markers associated with antimicrobial resistance directly from urine samples using next generation sequencing. *Sex. Transm. Infect.* **93**, 65–67 (2017).

114. Kumar, J. *et al.* An Improved Methodology to Overcome Key Issues in Human Fecal Metagenomic DNA Extraction. *Genomics Proteomics Bioinformatics* **14**, 371–378 (2016).

115. Giallonardo, F. D. *et al.* Full-length haplotype reconstruction to infer the structure of heterogeneous virus populations. *Nucleic Acids Res.* **42**, e115 (2014).

116. Ruppé, E. *et al.* Clinical metagenomics of bone and joint infections: a proof of concept study. *Sci. Rep.* **7**, 7718 (2017).

117. Kolb, M. *et al.* Next-Generation Sequencing for the Diagnosis of Challenging Culture-Negative Endocarditis. *Frontiers in Medicine* **6**, (2019).

118. Choutko, V. *et al.* Rare Case of Community-Acquired Endocarditis Caused by Neisseria
meningitidis Assessed by Clinical Metagenomics. *Frontiers in Cardiovascular Medicine* 6, (2019).

119. Tschumi, F. *et al.* Meningitis and epididymitis caused by Toscana virus infection imported to Switzerland diagnosed by metagenomic sequencing: a case report. *BMC Infect. Dis.* 19, 591 (2019).

120. Cobián Güemes, A. G. *et al.* Cystic Fibrosis Rapid Response: Translating Multi-omics Data into Clinically Relevant Information. *MBio* 10, (2019).

121. Mu, A. *et al.* Reconstruction of the Genomes of Drug-Resistant Pathogens for Outbreak Investigation through Metagenomic Sequencing. *mSphere* 4, (2019).

122. Lazarevic, V. *et al.* When Bacterial Culture Fails, Metagenomics Can Help: A Case of Chronic Hepatic Brucelloma Assessed by Next-Generation Sequencing. *Front. Microbiol.* 9, 1566 (2018).

123. Leo, S. *et al.* Detection of Bacterial Pathogens from Broncho-Alveolar Lavage by Next-Generation Sequencing. *Int. J. Mol. Sci.* 18, (2017).

124. Wilson, M. R. *et al.* Clinical Metagenomic Sequencing for Diagnosis of Meningitis and Encephalitis. *N. Engl. J. Med.* 380, 2327–2340 (2019).

125. Charalampous, T. *et al.* Rapid Diagnosis of Lower Respiratory Infection using Nanopore-based Clinical Metagenomics. doi:10.1101/387548

126. Thoendel, M. J. *et al.* Identification of Prosthetic Joint Infection Pathogens Using a Shotgun Metagenomics Approach. *Clin. Infect. Dis.* 67, 1333–1338 (2018).

127. Afshinnekoo, E. *et al.* Precision Metagenomics: Rapid Metagenomic Analyses for Infectious Disease Diagnostics and Public Health Surveillance. *J. Biomol. Tech.* 28, 40–45
(2017).

128. Gauffin Cano, P., Santacruz, A., Moya, Á. & Sanz, Y. Bacteroides uniformis CECT 7771 ameliorates metabolic and immunological dysfunction in mice with high-fat-diet induced obesity. *PLoS One* **7**, e41079 (2012).

129. Karlsson, F. H. *et al.* Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* **498**, 99–103 (2013).

130. Zeevi, D. *et al.* Personalized Nutrition by Prediction of Glycemic Responses. *Cell* **163**, 1079–1094 (2015).

131. Rubin, T. A., Gessert, C. E., Aas, J. & Bakken, J. S. Fecal microbiome transplantation for recurrent *Clostridium difficile* infection: report on a case series. *Anaerobe* **19**, 22–26 (2013).

132. Shahinas, D. *et al.* Toward an understanding of changes in diversity associated with fecal microbiome transplantation based on 16S rRNA gene deep sequencing. *MBio* **3**, (2012).

133. Bakken, J. S. *et al.* Treating *Clostridium difficile* infection with fecal microbiota transplantation. *Clin. Gastroenterol. Hepatol.* **9**, 1044–1049 (2011).

134. Chitnis, A. S. *et al.* Epidemiology of community-associated *Clostridium difficile* infection, 2009 through 2011. *JAMA Intern. Med.* **173**, 1359–1367 (2013).

135. Hourigan, S. K., Oliva-Hemker, M. & Hutfless, S. The prevalence of *Clostridium difficile* infection in pediatric and adult patients with inflammatory bowel disease. *Dig. Dis. Sci.* **59**, 2222–2227 (2014).

136. Sammons, J. S. & Toltzis, P. Recent trends in the epidemiology and treatment of *C. difficile* infection in children. *Curr. Opin. Pediatr.* **25**, 116–121 (2013).
137. Steiner C, B. & M, W. *HCUP projections: Clostridium difficile hospitalizations 2001 to 2013.* (Rockville, MD: Agency for Healthcare Research and Quality, 2014).

138. McFarland, L. V. *et al.* Recurrent Clostridium difficile disease: epidemiology and clinical characteristics. *Infect. Control Hosp. Epidemiol.* **20**, 43–50 (1999).

139. Fekety, R. *et al.* Recurrent Clostridium difficile Diarrhea: Characteristics of and Risk Factors for Patients Enrolled in a Prospective, Randomized, Double-Blinded Trial. *Clinical Infectious Diseases* **24**, 324–333 (1997).

140. Website. Available at: https://www.fda.gov/vaccines-blood-biologics/safety-availability-biologics/important-safetyalert-regarding-use-fecal-microbiota-transplantation-and-risk-serious-adverse. (Accessed: 29th October 2019)

141. Hourigan, S. K. *et al.* Fecal Transplant in Children With Clostridioides difficile Gives Sustained Reduction in Antimicrobial Resistance and Potential Pathogen Burden. *Open Forum Infectious Diseases* **6**, (2019).

142. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).

143. Hourigan, S. K. *et al.* Microbiome changes associated with sustained eradication of Clostridium difficile after single faecal microbiota transplantation in children with and without inflammatory bowel disease. *Aliment. Pharmacol. Ther.* **42**, 741–752 (2015).

144. Cohen, N. A. & Maharshak, N. Novel Indications for Fecal Microbial Transplantation: Update and Review of the Literature. *Dig. Dis. Sci.* **62**, 1131–1145 (2017).

145. Borody, T., Campbell, J., Torres, M., Nowak, A. & Leis, S. Reversal of Idiopathic
Thrombocytopenic Purpura [ITP] with Fecal Microbiota Transplantation [FMT]. *American Journal of Gastroenterology* **106**, S352 (2011).

146. Depommier, C. *et al.* Supplementation with Akkermansia muciniphila in overweight and obese human volunteers: a proof-of-concept exploratory study. *Nature Medicine* (2019). doi:10.1038/s41591-019-0495-2

147. Ivanov, A. Barriers to the Introduction of New Medical Diagnostic Tests. *Laboratory Medicine* **44**, e132–e136 (2013).

148. Miller, S. *et al.* Laboratory validation of a clinical metagenomic sequencing assay for pathogen detection in cerebrospinal fluid. *Genome Res.* **29**, 831–842 (2019).

149. Blauwkamp, T. A. *et al.* Analytical and clinical validation of a microbial cell-free DNA sequencing test for infectious disease. *Nat Microbiol* **4**, 663–674 (2019).

150. Schlaberg, R. *et al.* Validation of Metagenomic Next-Generation Sequencing Tests for Universal Pathogen Detection. *Arch. Pathol. Lab. Med.* **141**, 776–786 (2017).

151. Liu, L. *et al.* Comparison of next-generation sequencing systems. *J. Biomed. Biotechnol.* **2012**, 251364 (2012).
Supplementary Materials

Table S1. Supplementary Note 1. Comparison of available sequencing platforms. Platform is the sequencing machine referenced. Company is the current producer of the sequencing platform. Error rate is the percentage of the base pairs that are incorrectly read during sequencing. Average read length is the average length, in base pairs, of the reads generated by the sequencing platform when fed DNA. Bases per run is the throughput, in gigabases, that the machine can output every time it process a sample of DNA. Time spent per run is the amount of time the sequencer takes to process a sample of DNA. Cost per Gb is the amount of money, in US currency, to produce one Gigabyte of sequencing data. Release year is the year the sequencing platform was released for scientific use. Average read length, Bases per run, and release year were pulled from ‘Developments in high throughput sequencing’\textsuperscript{21}. For ABI Sanger, Error Rate and Cost per Gb was acquired from a paper entitled ‘Comparison of Next-Generation Sequencing Systems’\textsuperscript{151}

| Platform | Company         | Error Rate | Average Read Length (bp) | Bases per run (gigabases) | Time spent per run (hrs) | Cost per Gb (US Dollars) | Release Year |
|----------|----------------|------------|--------------------------|---------------------------|--------------------------|---------------------------|---------------|
| ABI Sanger | Applied Biosystems | .001       | 800                      | 0.0000768                 | 0.5-2.5                  | $5,000,000                | 2002          |
| Ion Torrent | Ion Torrent | ~1%        | 400                      | 2                          | 7.3                      | $460                      | 2013          |
| Platform | Company   | History | Accuracy | Cycle | Reads | Cost  | Year |
|----------|-----------|---------|----------|-------|-------|-------|------|
| MiSeq    | Illumina | 2013    | 0.2%     | 300   | 15    | 21-56 | $110 |
| Illumina | Solexa,   |         |          |       |       |       |      |
| HiSeq 2500 | Illumina | 2014    | 0.2%     | 250   | 300   | ~40   | $45  |
| PacBio RSII | Pacific Biosciences | 2013 | ~13%    | 13,500 | 12   | 2     | $600 |
| MinION   | Nanopore | 2015    | 32%      | 9,545 | 42    | ~0-50 | $750 |