Gastroenterology 2014;146:1437–1448

BASIC CONCEPTS IN THE MAMMALIAN GUT MICROBIOME

Meta'omic Analytic Techniques for Studying the Intestinal Microbiome

Xochitl C. Morgan[1,2] Curtis Huttenhower[1,2]

[1]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts; [2]The Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts

Nucleotide sequencing has become increasingly common and affordable, and is now a vital tool for studies of the human microbiome. Comprehensive microbial community surveys such as MetaHit and the Human Microbiome Project have described the composition and molecular functional profile of the healthy (normal) intestinal microbiome. This knowledge will increase our ability to analyze host and microbial DNA (genome) and RNA (transcriptome) sequences. Bioinformatic and statistical tools then can be used to identify dysbioses that might cause disease, and potential treatments. Analyses that identify perturbations in specific molecules can leverage thousands of culture-based isolate genomes to contextualize culture-independent sequences, or may integrate sequence data with whole-community functional assays such as metaproteomic or metabolomic analyses. We review the state of available systems-level models for studies of the intestinal microbiome, along with analytic techniques and tools that can be used to determine its functional capabilities in healthy and unhealthy individuals.

Keywords: Microbiome; Metagenomics; Metatranscriptomics; Tools.

The human microbiome comprises approximately 10^{14} bacteria and archaea, as well as fungi and viruses, which comprise roughly one kilogram of the average adult's body weight.[1] Our ability to study this remarkable system has changed dramatically in recent years—the cost of sequencing 1 million nucleotide bases of DNA decreased from $10,000 to $0.10 US dollars between 2001 and 2011 as a result of next-generation sequencing techniques.[2] This decrease in cost enabled large-scale surveys of human microbial diversity and function such as MetaHit[3] and the Human Microbiome Project,[4] which would have been prohibitively expensive as few as 15 years ago. These projects defined microbial compositions and their biomolecular functions in large populations. For the intestinal microbiota in particular, they showed hundreds of species, thousands of strains, and millions of bacterial genes.[3,4] Importantly, they also spurred rapid growth in analytic techniques, education programs, and numbers of trainees equipped to tackle complex sequence data from microbial communities.

Now that these studies have surveyed the wide degree of baseline microbial variation and sequenced thousands of human-associated microbe reference strains, the field is in a unique and unprecedented position to conduct well-informed investigations of the contribution of microbes to human health. Meta'omic studies may use 1 technique or combine several to address a multitude of questions, including but not limited to: whether changes in the composition of the microbiome are associated with, precede, follow, or cause the onset of disease; which microbial biochemical functions change in disease at the DNA, RNA, protein, and metabolite levels; how metabolic processes change in disease; and how interventions affect the composition and biomolecular function of the microbial community.

Shotgun-sequencing techniques, which include both DNA-focused metagenomic and RNA-focused metatranscriptome analyses, are especially useful in integrating microbial membership with biomolecular potential and activity in the human intestine. The low cost of nucleotide-based approaches make them an important component of any high-throughput experimental toolbox. Shotgun sequencing, in particular, provides strain- and gene-level information that is difficult to obtain from other technologies. The rapid evolution of sequencing technologies has been paralleled by corresponding increases in the availability and diversity of analytic tools and pipelines, and these have been

Abbreviations used in this paper: **IBD, inflammatory bowel disease; PCR, polymerase chain reaction; rRNA, ribosomal RNA; tRNA, transfer RNA; WMS, whole-metagenome or metatranscriptome sequencing.**

0016-5085/$36.00
http://dx.doi.org/10.1053/j.gastro.2014.01.049

followed by increased use of metaproteomic and metabolomic data. We review a subset of current analytic approaches for the rapidly evolving new field of meta'omics, as well as benefits and pitfalls of these techniques.

From Culture to 16S Ribosomal RNA Genes and Beyond

Culture of an organism has long provided, and still provides, one of the most detailed environments for study. However, most intestinal microbes are anaerobic and therefore are difficult to culture. In the mid-1970s, Woese et al[5] noted that portions of the gene encoding the small subunit 16S ribosomal RNA (rRNA) (hereafter referred to as 16S) were highly conserved among bacteria. Other internal regions of the gene are highly variable, possessing almost entirely unique sequences in most bacterial clades. Polymerase chain reaction (PCR) amplification, with universal primers, of a pool of DNA from a community of microbes, followed by cloning and sequencing, provides marker genes (a molecular nametag) that can be used to quantify bacterial taxa present within a sample. The Sanger sequencing method originally applied to this system has progressed through 454 technologies[6,7] to Illumina (San Diego, CA) sequencing,[8,9] which can sequence tens of thousands of 16S genes from a single sample with unprecedented affordability.

Methods for analyzing 16S sequencing data from the human microbiome and other environments are now well developed and have been reviewed elsewhere.[10–14] The most common approaches rely on microbial ecologic techniques, such as diversity analysis and ordination[15,16]; the potential imprecisions of 16S-based microbial classification and the associated sequence data have made it difficult to precisely identify microbes that might cause specific diseases, although this is improving.[11,12,17] Increasingly, 16S-based analyses of intestinal microbiomes of patients with specific disorders rely on reproducible diagnostic and prognostic biomarker discovery methods, which were developed from molecular epidemiology approaches in gene expression and genetic studies.[18–20]

Although 16S sequencing is the most widely used platform for studies of the gut microbiome because of its low cost (Table 1), it has several notable limitations. First, its accuracy depends on whether the observed proportions of 16S gene sequences reflect the proportion of bacteria in the sample, but the 16S gene is subject to copy number variation, as well as PCR primer and amplification bias. These biases range from being relatively minor to preventing detection of entire species or clades.[21,22] Furthermore, although 16S sequencing provides information about the overall composition of the microbial community, it does not provide information about the genomes of its members or their functions. For communities with many available reference genomes, such as the human intestine, it is possible to infer an approximate metagenome using methods such as PICRUSt.[23] These methods couple functions of gene products encoded by the most closely related sequenced genomes with observed taxonomic profiles to produce a functional profile.

Most importantly, 16S sequencing identifies only bacterial components of a community—not other types of microbes. However, amplicon approaches, which incorporate the 18S ribosomal subunit gene and internal transcribed spacer sequences for analysis of eukaryotes, are rapidly improving.[24,25] Now that the cost difference has narrowed between 16S and whole-metagenome or metatranscriptome sequencing (WMS) (Table 1), it is becoming feasible to design 2-stage experiments that incorporate both approaches[26] in a cost-effective manner, or even exclude 16S analysis.

Table 1. Microbiome Sequencing Approaches: Costs and Caveats

Microbial data of interest	Relevant technologies	Current approximate cost (US$/sample)	Notes
Bacterial/archaeal composition	16S on stool or mucosa	$10–$100	Primer bias; see recommendations such as 515F/806R[96]
Fungal composition	Internal transcribed spacer analysis of stool or mucosa	$10–$100	Fungi frequently are difficult to lyse; use of a bead-beater and optimization of DNA extraction protocols is recommended[97,98]
Bacterial/archaeal/viral/fungal composition and function	WMS on stool	$100–$500	WMS on mucosa will have extensive host contamination, but WMS on stool typically contains <1% host reads
Viral composition and function	WMS on stool DNA or RNA prepared from complementary DNA, after enrichment for viral fraction	$100–$500	Viral DNA can range from <0.1% to a significant fraction of stool samples; enrichment before sequencing can improve detection consistency[29,99]
Bacterial/archaeal/viral/fungal transcriptome	WMS on stool complementary DNA prepared from RNA	$200–$1000	Mucosa will have extensive host material and low bacterial biomass is likely; stool is substantially technically simpler; depletion of rRNA is crucial
Host genomics, transcriptomics, proteomics, or immunology	Sample from mucosa	Varies	Mucosal biopsy specimens can be paired with stool or multiplexed to joint host and microbial assays for parallel multi'omic data