Comparing 16S rRNA Marker Gene and Shotgun Metagenomics Datasets in the American Gut Project Using State of the Art Tools

Embriette R. Hyde, Jon Sanders, Anupriya Tripathi, Qiyun Zhu, Rob Knight
American Gut Project and University of California, San Diego

What is the American Gut Project?
The American Gut Project is the country’s largest open source citizen science project in existence today. The project operates on the basis of crowdfunding—which means that project participants contribute both samples and funds to process those samples. This wouldn’t have been possible even as recently as 5 years ago—but the exponential decrease in the cost of sequencing has enabled regular people to contribute to meaningful science. By collecting samples and detailed health, lifestyle, and dietary details from participants, we are building a database that can be leveraged to identify unhealthy and healthy microbiome states and to move humans from bad to good states.

Since the Human Microbiome Project (1, 2), we’ve learned from hundreds of studies that the microbiome - all of the bacteria, archaea, viruses, and fungi that live on us and in us - is associated with a number of disease states from inflammatory bowel disease (IBD) to obesity to autism, and that the microbiome helps keep us healthy by encouraging proper immune system development and function, by aiding digestion, and by producing essential vitamins that human cells can’t. This means that we might one day be able to target the microbiome as a treatment option for disease; however, in order to move to clinical trials and eventually into clinical practice, we need to learn more. Statistical power is key, and therefore, so are large sample sizes. The American Gut Project (americangut.org) is making great strides toward this goal. As of early 2017, the project has raised over $2 million and processed over 14,000 samples from over 10,000 individuals. But we still need more samples, and we’re building the cohort further by reaching out to specific groups on both ends of the health spectrum.

American Gut utilizes benchmarked, open protocols available through the Earth Microbiome Project (earthmicrobiome.org) facilitating meta-analyses using the American Gut data together with any other microbiome dataset that has been processed using these protocols. You may be wondering how feasible such comparisons are, particularly when handling thousands of samples. Developers in the Knight Lab have designed a sophisticated set of analysis tools to make such analyses relatively straightforward. One of these tools is called Qiita, which will be described more below.

What is 16S rRNA marker gene analysis?
The 16S rRNA gene is a ribosomal gene thought to exist in all Bacteria and Archaea. The gene is characterized by conserved and highly variable regions; primers targeting conserved regions that flank variable regions amplify the variable regions, which enables taxonomic classification down to the genus level. Many microbiome studies today target the V4 region of the 16S rRNA gene as the primers that amplify this portion are least susceptible to bias and are able to yield the greatest amount of information taxonomically from most samples types. Molecular barcoding techniques permit the sequencing of hundreds of samples at a time on single Illumina MiSeq or HiSeq runs, dropping the cost for sequencing dramatically. Because of this, most microbiome studies to date have utilized the 16S rRNA marker gene profiling approach, which yields information about which
taxonomic groups (from phylum to genus level) are present in a sample and at what relative abundance. Most American Gut samples have been processed in this way as the low cost makes personal microbiome sequencing attainable for many.

**What is shotgun metagenomics?**

Marker gene analysis surveys only a small portion of each genome. In contrast, shotgun metagenomics surveys the entire genomes of all organisms present. A principal advantage of shotgun metagenomics is that it is much less susceptible to the biases inherent in target gene amplification: the primers typically used for 16S analysis don’t match eukaryotic rRNA genes, and can’t give information about viruses, which don’t have rRNA genes. By sequencing all DNA present, shotgun metagenomics yields information about the entire community including viruses and fungi. Because most regions of the genome are more highly variable than are rRNA genes, shotgun data can also give much higher resolution, distinguishing between closely related organisms. With sufficient sequencing depth, shotgun metagenomic data can also yield direct information about the presence and abundance of particular functional gene pathways, or be assembled into draft genomes, yielding insights into the physiological capabilities of abundant organisms in the community. Historically, this has required much deeper sequencing than that required by marker gene surveys; however, the Knight lab has been working on protocols to lower the cost of shotgun metagenomics as well, and has developed a shallow shotgun metagenomics protocol that will reveal information about microbial species and functional gene pathways present in stool samples. Your samples have been processed for both marker gene and shotgun metagenomics sequencing and analysis.

**What is Qiita?**

In the Knight Lab, we utilize several open source tools for analyzing large microbiome datasets (both marker gene and shotgun metagenomics) to obtain meaningful associations between the microbiome and health, disease, and lifestyle factors. One of those tools is called Qiita. Qiita is a web service developed and hosted in the Knight Lab with the goal of standardizing the data representation of multi-omics datasets (sequencing, proteomics, transcriptomics, and metabolomics) to improve our ability to understand the community composition and function of the microbial world at high resolution. Qiita’s design allows users to keep track of the vast amount of data generated by current state of the art multi-omics technologies, simplifying the integration of multiple datasets to create meta-analyses, empowering the user with new approaches and points of view that facilitate the discovery of new insights in their data. The Qiita main server, accessible at qiita.microbio.me, provides database and compute resources to the global community, alleviating the technical burdens that are typically limiting for microbial ecology researchers. While Qiita has focused to date on analysis of marker gene data, we have been adding capabilities for basic analyses of other ‘omics tools, including shallow shotgun sequence profiling and visualization of small-molecule metabolomics.

**Tools for Analyzing Shotgun Metagenomics Data**

The identification of taxonomic and functional compositions of a metagenome (a.k.a., profiling) is typically achieved by mapping the sequencing data against a database of reference genomes and genes. A variety of computational tools are available for this purpose. Some tools, such as MetaPhlAn (3) and mOTU (4), rely on the mapping of a selected list of marker genes extracted from
the metagenome; while other tools, such as Kraken (5) and Centrifuge (6), can take advantage of the whole metagenome. The latter category typically achieve high sensitivity and resolution (species and even strain level), at the cost of increased computational expense. This concern is not trivial when analyzing large amount of sequencing data, which is common in modern studies. As a workaround, researchers usually have to use “heuristic” algorithms, which surveys only a subset of the database to get quick, but suboptimal matches (meaning chances of false assignment). Our collaborator, Dr. Dan Knights at the University of Minnesota, developed SHOGUN (https://github.com/knights-lab/SHOGUN), a highly efficient and accurate whole metagenome profiling tool. It features an “exhaustive” yet rapid alignment algorithm, which guarantees the finding of globally best matches. Our benchmarking results have confirmed the remarkable performance of SHOGUN on real shotgun metagenomic datasets.

A Cross-Comparison of 16S and Shotgun Metagenomics Data: The AGA Cohort

A few months ago you were sent American Gut kits and invited to join our group of over 10,000 citizen scientists. We’ve performed a series of analyses on these samples to illustrate the similarities and differences between 16S rRNA marker gene sequencing and shotgun metagenomics. First, we’ll show you where you fit on the greater American Gut Project “map” (Figure 1).

Figure 1. An unweighted UniFrac-based principal coordinates plot shows the AGA cohort samples (large blue dots; n = 53) in the context of all American Gut Project stool samples (small red dots; n = 11,887).
Looking at taxa summaries from the marker gene sequencing effort (Figure 2), we see that like most human stool samples, the major phyla present are Bacteroidetes and Firmicutes. Three individuals have notable abundances (near 50% to over 70%) of Proteobacteria; in comparison, the average amount of Proteobacteria observed across the entire American Gut population is less than 10%. While Gammaproteobacteria are well recognized to be increased in a number of diseases, including IBD, in the case of the AGA cohort individuals, the main taxa are of the alpha and betaproteobacteria classes. In participants 000069756 and 000070851, the main genus is *Brevundimonas* (Alphaproteobacteria), followed by *Rhodospirillum* (Betaproteobacteria) and *Comamonas* (Betaproteobacteria), respectively. The major genus present in participant 000069705 was *Sphingobium* (Alphaproteobacteria), followed by *Sphingopyxis* (Alphaproteobacteria) an unclassified genus in the Xanthomonadaceae family (Gammaproteobacteria).

![Phylum Level Taxonomy Summaries (16S) of 53 AGA Members](image)

*Figure 2.* Stacked bar charts illustrate the mean relative abundance of bacterial phyla detected in AGA cohort samples using the 16S marker gene sequencing protocol.
Looking at taxa summaries from the shotgun metagenomics effort, we see that in general, there is some consistency between the 16S and shotgun metagenomics approaches although some obvious differences are noted. Two of the three samples with high levels of Proteobacteria via amplicon sequencing also have high levels with shotgun metagenomics sequencing. The most obvious difference between the two approaches is the increased abundance of Bacteroidetes (and corresponding lower abundance of Firmicutes) overall in the shotgun metagenomics cohort. Additionally, while most of the taxa detected were bacteria, there were also Archaea, Fungi (eukaryotes), and Viruses detected (Figure 3). There were a total of 5 archaeral genera, one parasitic (eukaryotic) genus, 7 fungal (eukaryotic) genera, and 25 viral genera detected in the dataset.

![Phylum Level Taxonomy Summaries (SHOGUN) of 53 AGA Members](image)

**Figure 3.** Phylum level stacked bar charts reveal the relative abundances of Archaea, Bacteria, Fungi, and Viruses present in the AGA cohort. Taxonomy information and relative abundances were obtained using SHOGUN. The vast majority of detected taxa were bacterial, though a few individuals had archaea, fungi, or viruses present at a relative abundance of 1% or more.

In most cases, non-bacterial taxa were at low abundance, or not detected; however, in a handful of individuals, non-bacterial taxa were detected at a relative abundance of 0.5% or more. **Table 1** lists the taxonomic classifications of all non-bacterial genera detected in the full AGA dataset of 53 samples. **Table 2** lists the six AGA samples which had at least 0.5% relative abundance of non-bacterial taxa present, with the relative abundances of bacterial, archaeal, fungal, and viral taxa listed.
Archaeal Genera

Halorubrum (Halobacteria);
Methanobrevibacter (Methanobacteria);
Methanosphaera (Methanobacteria);
unclassified Thermoplasmata genus;
Methanomassiliicoccus (Thermoplasmata)

Parasitic (Eukaryotic) Genera

Theileria (Apicomplexa)

Fungal (Eukaryotic) Genera

Candida (Saccharomycetes);
Eremothecium (Saccharomycetes);
Kluuyveromyces (Saccharomycetes);
Saccharomyces (Saccharomycetes);
Sugiyamaella (Saccharomycetes);
Thermothelemyces (Sordariomycetes);
Thielavia (Sordariomycetes)

Viral Genera

Unclassified viral genus;
unclassified Microviridae genus;
Phix174microvirus (Microviridae);
Alphapapillomavirus (Papillomaviridae);
unclassified Retroviridae genus;
unclassified Myoviridae genus;
Felixo1virus (Myoviridae);
Hp1virus (Myoviridae);
P1virus (Myoviridae);
P2 virus (Myoviridae);
unclassified Podoviridae virus;
Epsilon15virus (Podoviridae);
F116virus (Podoviridae);
P22virus (Podoviridae);
T7virus (Podoviridae);
unclassified Siphoviridae genus;
C2virus (Siphoviridae);
D3virus (Siphoviridae);
Lambdavirus (Siphoviridae);
Nonagvirus (Siphoviridae);
Phietavirus (Siphoviridae);
Phifelvirus (Siphoviridae);
Sfi11virus (Siphoviridae);
Sfi21dt1virus (Siphoviridae);
Sk1virus (Siphoviridae)

Table 1. List of non-bacterial (archaeal, eukaryotic, and viral) taxa detected in the AGA cohort.

<table>
<thead>
<tr>
<th>Barcode</th>
<th>000069766</th>
<th>000069704</th>
<th>000069770</th>
<th>000069707</th>
<th>000069751</th>
<th>000069719</th>
</tr>
</thead>
<tbody>
<tr>
<td>Viruses</td>
<td>0.01%</td>
<td>0.001%</td>
<td>0.006%</td>
<td>0.0003%</td>
<td>1.59%</td>
<td>23.3%</td>
</tr>
<tr>
<td>Eukaryota</td>
<td>0.004</td>
<td>0.0004%</td>
<td>1.42%</td>
<td>0%</td>
<td>0%</td>
<td>0.008%</td>
</tr>
<tr>
<td>Archaea</td>
<td>0.79%</td>
<td>0.90%</td>
<td>0%</td>
<td>1.36%</td>
<td>0.14%</td>
<td>0%</td>
</tr>
<tr>
<td>Bacteria</td>
<td>99.19%</td>
<td>99.09%</td>
<td>98.57%</td>
<td>98.64%</td>
<td>98.27%</td>
<td>76.69%</td>
</tr>
</tbody>
</table>

Table 2. Relative abundances of viral, eukaryotic, archaeal, and bacterial taxa are listed for the six AGA samples in which at least one non-bacterial micro-organism was detected at a relative abundance of at least 0.5%.
In participant 000069766, the two dominant archaea were *Methanobrevibacter* and *Methanosphaera*. In participants 000069704 and 000069707, these were *Methanomassiliicoccus* and *Methanobrevibacter*. In participant 000069770, the major eukaryotic taxa was *Candida*. In participant 000069751, the major viral genus was Nonagivirus. The major bacterial taxa in participant 000069719 were *Pseudomonas_phage_O4* and *Pseudomonas_phage_PA11*.

A Procrustes analysis (least squares orthogonal mapping) was performed to compare the taxonomic patterns generated from 16S rRNA sequencing and shotgun metagenomic sequencing. This analysis stretches, rotates and superimposes sample coordinates obtained from one dataset over the other, while preserving distances within each individual matrix. The goal is to find the best fit between two matrices to infer whether one dataset coherently captures the properties of the other. **Figure 4** below shows that different level of information is captured by 16S rRNA sequencing and shotgun metagenomic sequencing based on sample clustering.

**Figure 4.** Procrustes between unweighted UniFrac based distances for 16S rRNA dataset (white line) and Bray-Curtis based distances for shotgun metagenomic dataset (maroon line). The coherence between the two datasets was scored by the goodness of fit value obtained from PROTEST ($M^2 = 0.584$; Mantel R-statistic=0.48260).
Assessing the top 20 pathways present across the dataset, we see a stable pattern in which most pathways are at similar relative abundances in all samples. However, as seen in Figure 5, in the samples in which Proteobacteria were present at relative abundances of over 70%, three pathways are noticeably higher: chorismate biosynthesis from 3-dehydroquinate, chorismate biosynthesis I, and peptidoglycan biosynthesis I. The potential significance of the increased abundance of these three pathways is unclear, particularly because chorismate biosynthesis is a well-described pathway in *Saccharomyces* and *Chlamydia*.

![Relative Abundances of the Top 20 Gene Pathways in the AGA Cohort](image)

**Figure 5.** The mean relative abundances (in relation to each other) of the top 20 pathways detected across all AGA cohort samples.

**Case Study: Deep Shotgun Metagenomic Sequencing of Longitudinal Stool Samples Collected from a Single IBD Patient**

As demonstrated by the analysis of your samples above, even relatively shallow shotgun sequencing offers advantages in taxonomic resolution and breadth over marker gene analyses. More deeply sequenced shotgun metagenomics datasets offer fundamentally different analytical capabilities. Assembling sequence reads into draft microbial genomes allows us to begin drawing links between predicted microbial functions and the specific organisms performing those functions, and to follow their distribution and evolution across samples.

Longitudinally sampled subjects offer especially rich datasets for deeper shotgun metagenomic analysis. Because we expect the same organisms to exist across multiple samples at changing frequencies, we can statistically group or “bin” genome fragments from the same underlying genome into draft microbial genomes, or “Metagenomically Assembled Genomes” (MAGs). These
MAGs allow for the detailed exploration of the functional capacity of individual lineages, and can form the basis of cross-omics studies linking gene content, transcription, proteomics, and small molecule biosynthesis.

Here, we will explore some of these concepts using a longitudinally-sampled fecal microbiome dataset from an individual with undiagnosed IBD-like symptoms. This dataset consists of samples collected monthly or semi-monthly between 2012 and 2015, with a subset of high-resolution daily sampling at the end of the series.

As we showed for the cross-sectional dataset above, basic taxonomic profiling of this time series using these whole-genome shotgun sequences offers some notable improvements over the 16S marker-gene survey data. The overall differences in relative abundance between these two datasets are likely the result of biases associated with amplification of the 16S gene in the marker survey, resulting in substantial under-representation of Bacteroidetes (teal bars) relative to the shotgun data. The decrease in bias and increase in taxonomic resolution afforded by the shotgun analysis also allows us to discern important trends that are not apparent in the 16S data: for example, the gradual replacement of Enterobacteriaceae (yellow bars) by Morganellaceae (pink bars) in the second half of the time series; as well as the transient bloom of Hafniaceae (black bars) towards the beginning of the series (Figure 6).

Figure 6. Taxonomic profiling of monthly or semi-monthly samples collected from an individual IBD patient between December 2011 and July 2015. Top panel, 16S profile; bottom panel, shotgun metagenomics profile.
The increased taxonomic resolution of these shotgun data also allows us to examine relationships over time between bacteria of the same genus. For example, we can plot the estimated relative abundances of nine *Bacteroides* species detected in the shotgun dataset over time, as well as their pairwise correlations (Figure 7). The two most abundant species, *Bacteroides cellulosilyticus* and *Bacteroides vulgatus*, are strongly anticorrelated in abundance, potentially reflective of niche competition. Notably, a third species (*Bacteroides fragilis*) is not strongly correlated to either, but shows peaks in abundance in the daily-sampled time-series that correspond to periods of transition between *B. cellulosilyticus* and *B. vulgatus* dominance.

![Figure 7](image)

**Figure 7.** Relative abundance of *Bacteroides* species over time in an individual IBD patient as detected by shotgun sequencing. Left panel, relative abundance of three selected *Bacteroides* species. *B. cellulosilyticus* and *B. vulgatus* are strongly anticorrelated, while *B. fragilis* spikes in abundance during periods of transition between the other two. Right panel, pairwise correlation plots for all *Bacteroides* species (plus time, bottom right). Above diagonal, correlation strength and direction indicated by colored circles (red is inverse correlation). Below diagonal, abundance of both species per pair plotted per timepoint.

Both of the above observations depend on the use of reference databases to classify sequence reads. This works well for many human-associated microbial communities, where most species are well-represented in existing databases. But for more undersampled communities and species, deep shotgun sequencing and assembly of MAGs allows for genome-resolved exploration across samples while growing the available databases to improve future studies.

We performed de novo metagenomic assembly on each of the time-series samples, using the pattern of differential representation across samples to bin the assembled reads into draft MAGs. For illustration, we show the results of the binning for one sample. In Figure 8, 131 candidate MAGs are represented by a circular dendrogram in the center, with MAGs showing similar abundance distributions clustered closer together on the dendrogram. The relative abundance of each of these MAGs across a portion of the monthly time series is represented for each of the samples by concentric circles, with earlier time points closer to the center. A darker shaded square indicates a higher relative abundance of that MAG at that timepoint.
Figure 8. Visualization of binned genomes (MAGs). Each bin is a leaf on the central dendrogram; each spoke depicts its abundance across a portion of the monthly timeseries. MAGs with similar abundance profiles are nearer one another on the dendrogram. Statistics per MAG (assembled length, GC-content, completion and redundancy as measured by expected universal single-copy genes) displayed as bar graphs around edges; information per stool sample, including taxonomic abundance profile and proportion of human DNA detected in sample, displayed in the bar graph at the 3 o’clock position. For context, the MAG depicted in more detail in Figure 9 (Bin 55) is highlighted at left.

We've visualized some statistics about the MAGs themselves around this plot, including total genome length (innermost circle), GC-content, completion (an estimate of how much of the genome is represented in the MAG, based on presence of universal marker genes), and redundancy (an estimate of over-binning based on redundant copies of those marker genes). From this sample alone, we recovered 56 genomes at more than 90% completeness.

We've also visualized some metadata about the samples, including an illustration of the taxonomic profile estimated by SHOGUN and the proportion of raw reads that mapped to a reference human genome, a proxy for host tissue in the stool sample. For this individual, months 7-12 of the time series had unusually high proportions of human DNA, potentially indicative of overall inflammation; this also corresponds to the period of Hafniaceae bloom noted above. Intriguingly, the return to normal levels of human DNA correlates with the extinction of several MAGs (at positions 11 to 12 o’clock) from the community.
Critically, this assembly-based analysis allows exploration of unknown bacteria — those that are currently not represented in databases, also known as 'microbial dark matter.' One such example in the illustrated assembly is MAG #55, highlighted near the 9 o’clock position in Figure 8. This genome was relatively abundant across the time series, but could not be classified below the phylum level (Proteobacteria), and would likely be missed by reference-dependent profiling.

We were able to assemble MAGs related to #55 out of a number of other samples, allowing us to perform pan-genomic analysis of gene content. We’ve illustrated several of these in Figure 9. Five MAGs with high phylogenetic similarity to MAG #55 are represented. A ‘core genome’ containing identical gene complement, including all expected single-copy universal marker genes, is present in each of the five MAGs. Other blocks of similar functional genes are present in between two to four of the MAGs, indicating likely regions of misassembly or gene exchange. Further investigation of the functional gene content of the core genome of MAG #55 will provide further insights into its potential role in the gut microbiome.

Figure 9. A representation of the pangenome of one MAG from the overall assembly, “Bin 55,” predicted to be from a lineage of Proteobacteria. Genome bins assembled independently from five separate samples are depicted. Individual predicted protein coding genes are represented as the leaves of the central dendrogram, with the set of genes found across all bins—the minimal ‘core genome’ for this bacterium—highlighted in the lower right. Other large clusters of proteins found across multiple bins (PCBs) are also highlighted.
References


Thank you for your participation in this project. De-identified results from the AGA cohort are available at ftp://ftp.microbio.me/aga-results. Your individual results can be accessed at the American Gut Project participant portal.

These data were presented at Digestive Disease Week® 2017 at the session titled “Microbiome Active Learning Session 200: The Non-Bacterial Gut Microbiome in Health and Disease.” Visit www.ddw.org for more information.

To learn more, visit the AGA Center for Gut Microbiome Research and Education at www.gastro.org/microbiome.