Exploring IBD in the Context of the American Gut Project Using State of the Art Tools

Jose A. Navas Molina, Embriette R. Hyde, Larry Smarr, William Sandborn, John Chang, Brigid Boland, Hongwei Zhou, Ye Chen, Jan Suchodulski, Yoshiki Vazquez Baeza, Zech Xu, Rob Knight

What is American Gut?

The American Gut Project [1] is the country's largest open source citizen science project in existence today. The project operates on the basis of crowdfunding-meaning project participants contribute both samples and funds to process those samples. This wouldn't have been possible even as recently as five 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 [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 the clinic, we need to learn more. Statistical power is key-and therefore, so are large sample sizes. The American Gut Project is making great strides toward this goal. As of May 2016, the project has raised over $1.1 million on Fundrazr and processed over 9,000 samples from over 7,000 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 (i.e. individuals with cancer and athletes).

American Gut utilizes benchmarked, open protocols available through the Earth Microbiome Project [3] 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, however, 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.

What is 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.

Using Meta-analyses to Look at the Microbiome in IBD

A few months ago you were sent American Gut kits and invited to join our large group of citizen scientists. We've performed a series of analyses to examine the microbiome of IBD. We combined the American Gut project data, including your own samples, with five other datasets: a Chinese IBD cohort (Hongwei Zhou and Ye Chen), a local IBD cohort (Bill Sandborn), a single individual diagnosed with Crohn's disease (CD) with longitudinal samples spanning months (Larry Smarr), and a previously published adolescent CD cohort [4]. We also combined these studies with a dog IBD study to assess species specific differences (or similarities). Leveraging Qiita, a meta-analysis including all of these studies together with the entire American Gut cohort and the AGA samples was done and is described here to illustrate not only how sophisticated tools can facilitate analysis of multiple datasets comprising billions of base pairs of sequencing data, but also to explore the microbiome in IBD (and in healthy subjects) across geographical space.

One of the main ways to compare microbial communities across samples is through beta diversity. We calculate beta diversity using a tool called UniFrac, creating a distance matrix comparing each sample in the dataset to each other sample in the dataset. UniFrac assigns a number between 0 and 1 based on the amount of shared branch length on the phylogenetic tree. The closer to 1, the more similar the microbial communities are; the closer to 0, the more dissimilar. As you can imagine, looking at such a distance matrix isn't very helpful when trying to visualize similarities and differences between microbial communities. Using principal coordinates analysis (multi-dimensional scaling) enables one to visualize the data present on the distance matrix. PCoA assigns samples a location along vectors/axes, based on how much variation between samples is attributed to each vector. For example, PC1 in **Figure 1** below is the vector responsible for the most variation in the samples-7.2%. Typically, PCoA plots are visualized in 2D or 3D (2 or 3 vectors).

Figure 1 below shows a principal coordinates analysis (PCoA) plot. Each dot on the plot represents a single sample; the closer together two samples are on the plot, the more similar the microbial communities represented by those dots-the further apart two samples are, the more dissimilar the microbial communities. The dots in the plot below have been colored by study. The smaller blue dots are samples from the American Gut project (includes both healthy and IBD individuals), while the red samples are the AGA cohort samples (your sample fits into this group). The Chinese (green), Sandborn (light blue), Smarr (yellow), Gevers (purple), and dog (orange) cohorts are also colored by study.

The first thing you may notice is that the AGA samples are widely distributed, as are the American Gut samples. Most AGA samples came from healthy individuals, i.e., those that do not suffer from IBD. Notably, the samples from the local Crohn's individual cluster completely separately from the local IBD samples (Sandborn cohort, mixture of CD and ulcerative colitis patients), the Gevers et al. 2014 adolescent CD samples, and the Chinese CD cohort. The Chinese cohort also forms an obvious cluster; this is likely due to differences in diet between the Chinese and (largely) North American (i.e. Westernized) cohorts. Interestingly, some samples from two of the human IBD cohorts cluster with the dog samples, indicating that microbially, these samples are more similar to dog samples than to some other human samples.

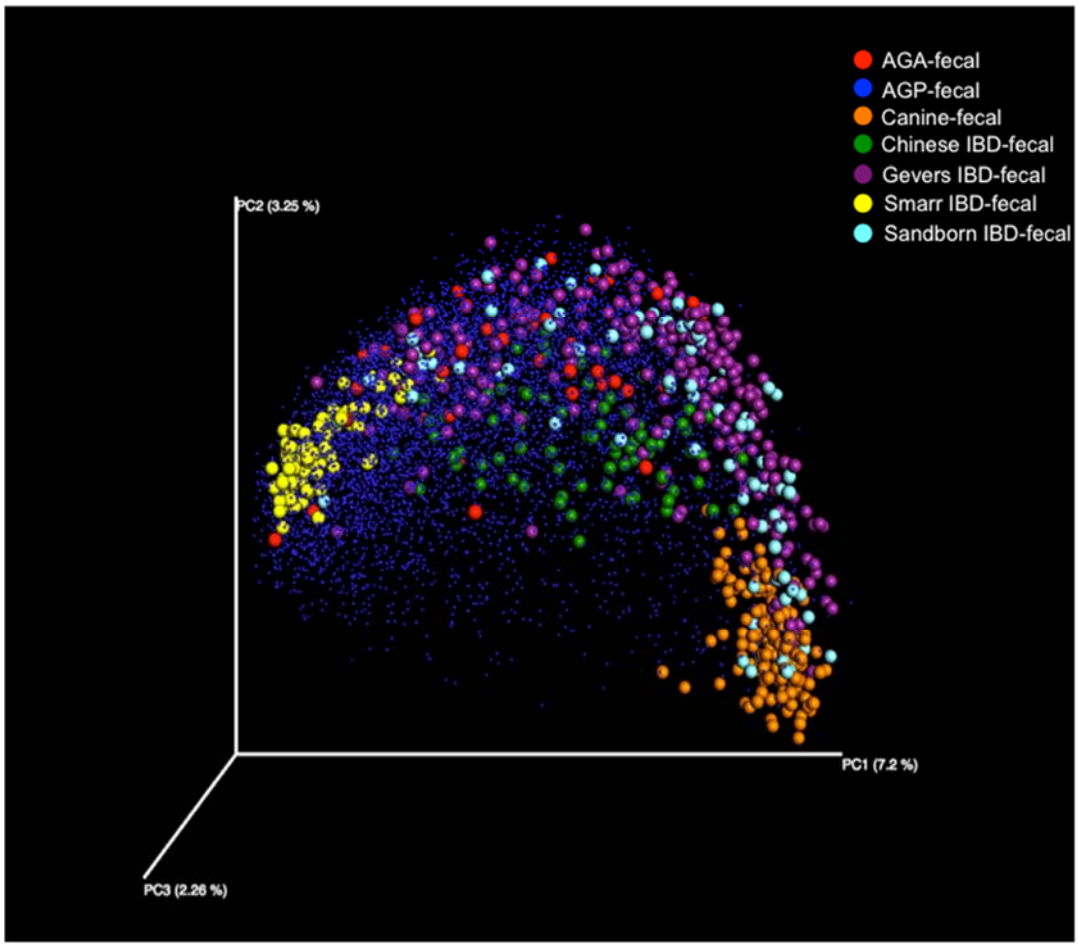


Figure 1. PCoA plot illustrating the seven datasets analyzed in the meta-analysis presented here. Each dot represents a single sample; dots are colored by study as indicated in the key.

Coloring by whether the study subject has IBD or not (**Figure 2, below**) or by disease subtype (**Figure 3, below**), reveals no obvious clustering patterns, indicating that overall microbiome composition is not notably associated with specific disease subtypes-at least not among the seven studies presented here. This may be due to the large size of the American Gut cohort (over 6,000 fecal samples) compared to the other studies presented here. Additionally, any specific associations with disease subtype could be present at levels not detectable with the methodology used here; i.e., at the individual species or even strain level. Notably, American Gut data about disease status are self-reported and do not reflect whether an individual is suffering from active disease or not, highlighting the potential benefits of collecting samples from well characterized diseased cohorts to fully piece together the connection between the microbiome and IBD.

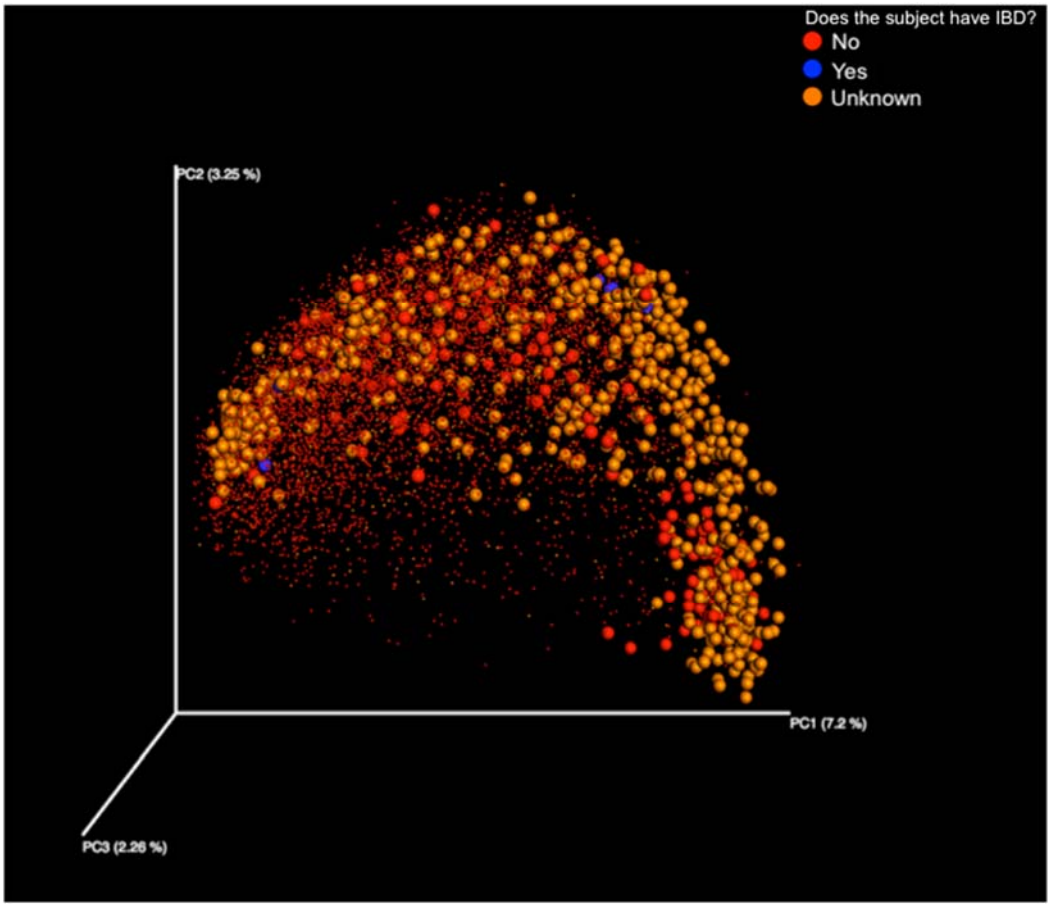


Figure 2. PCoA plot with samples colored according to IBD status (yes, no, or unknown).

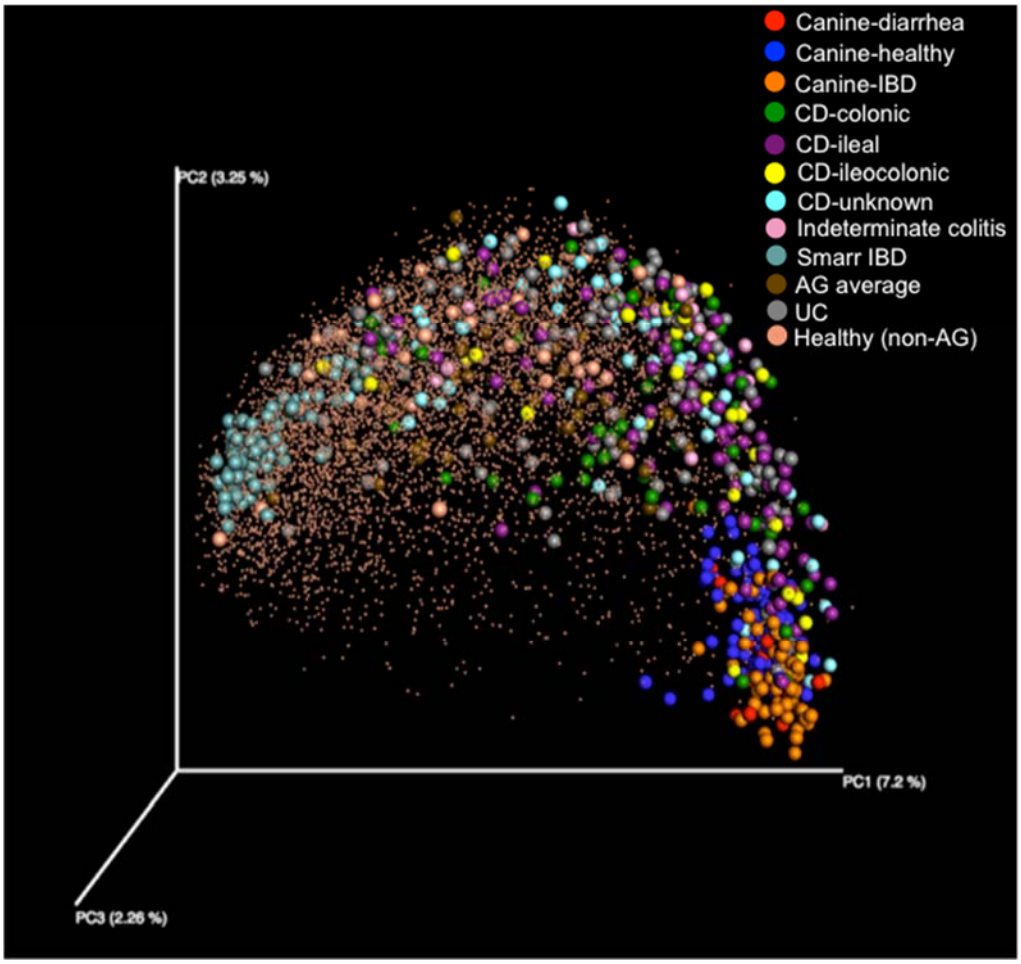


Figure 3. PCoA plot illustrating samples colored by disease subtype (canine health, canine diarrhea, canine IBD, CD-colonic-CD-ileal, CD-ileocolonic, indeterminate colitis, Smarr IBD, AG average, UC, healthy (non-AG cohort).

We can next take a look at bar charts that illustrate the relative abundance of bacterial taxa present in these sample groups. Shown below (**Figure 4)** is a phylum level plot. Each of the three domains of life – Bacteria, Archaea, and Eukarya – are further divided into subclassifications: phylum (most broad), class, order, family, genus, and species (most specific). There is some variation to the schema depending on the organism (i.e., there may be sub-orders or sub-species), but here we will refer to the 6 sub-divisions listed above). To give an idea of the range represented by these classifications, there are around 50 recognized bacterial phyla, there thousands of recognized bacterial genera, and tens of thousands of bacterial species! In the plot below, we see that the phylum Verrucomicrobia is present in every human cohort but the Chinese cohort, which has a small abundance of Fusobacteria. Verrocomicrobia are represented by one genus, *Akkermansia*, which is called mucolytic due to its activity in breaking down mucus in the human digestive tract. Canine samples are also largely different from human samples-with the community dominated by Firmicutes and very little Bacteroidetes, as well as a small amount of Verrucomicrobia.

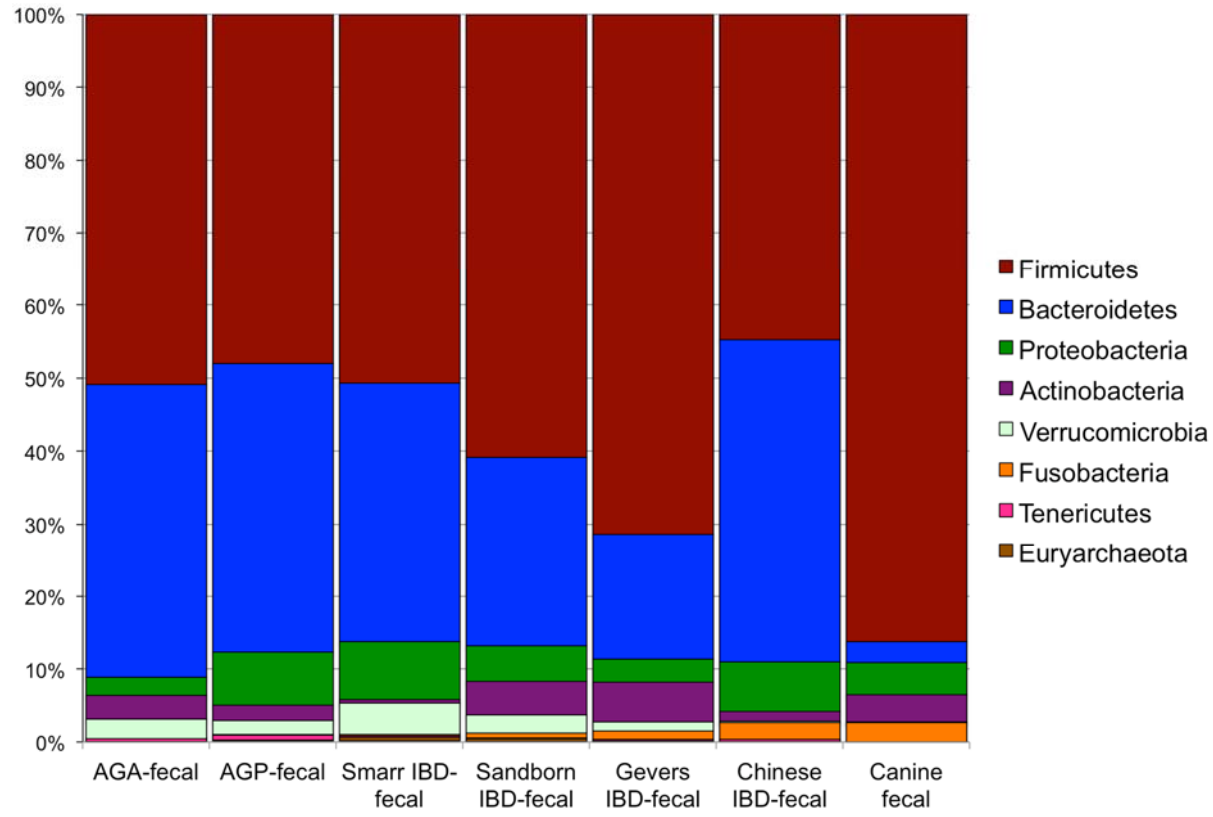


Figure 4. The mean relative abundances of bacterial phyla present in each cohort used in this meta-analysis is illustrated.

We can also look at disease subtype, as seen in **Figure 5** below. No notable differences are observed at the phylum level; however, Verrucomicrobia are only detected in ileal and ileocolonic CD but not in colonic CD.

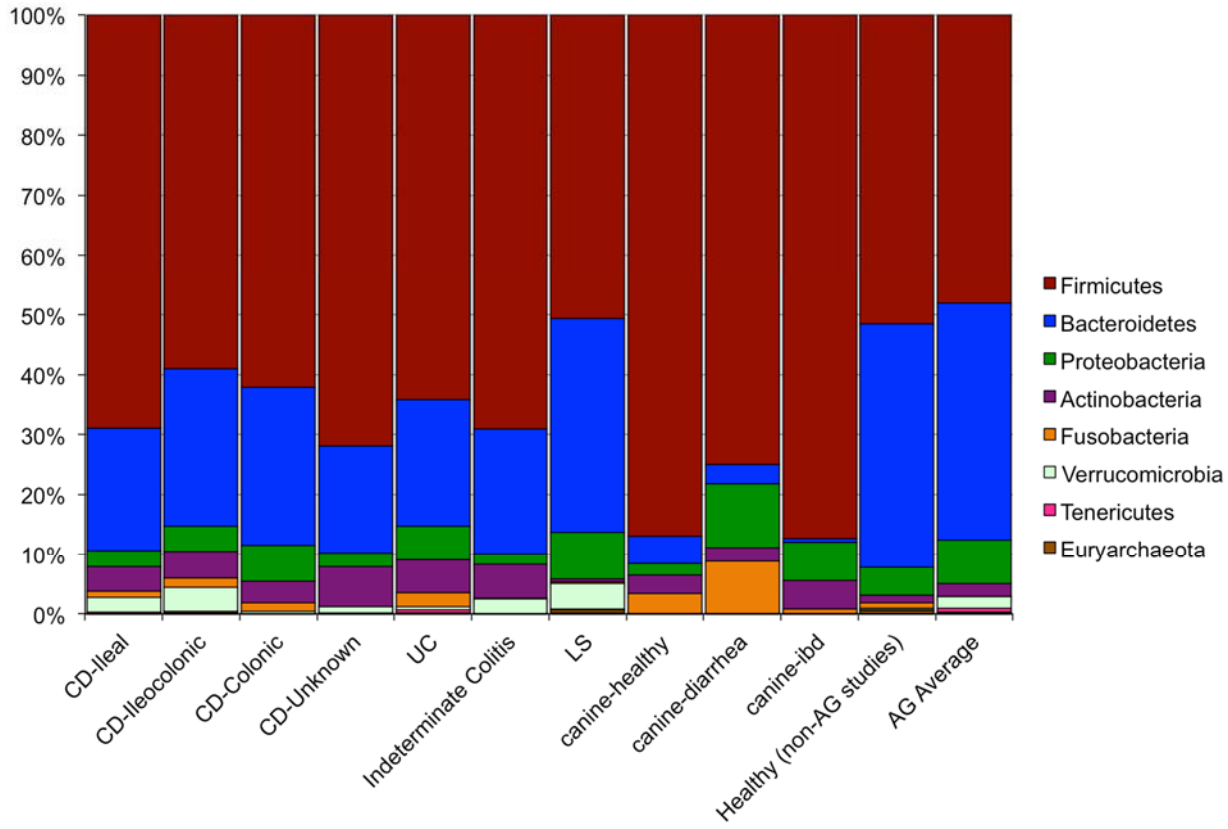


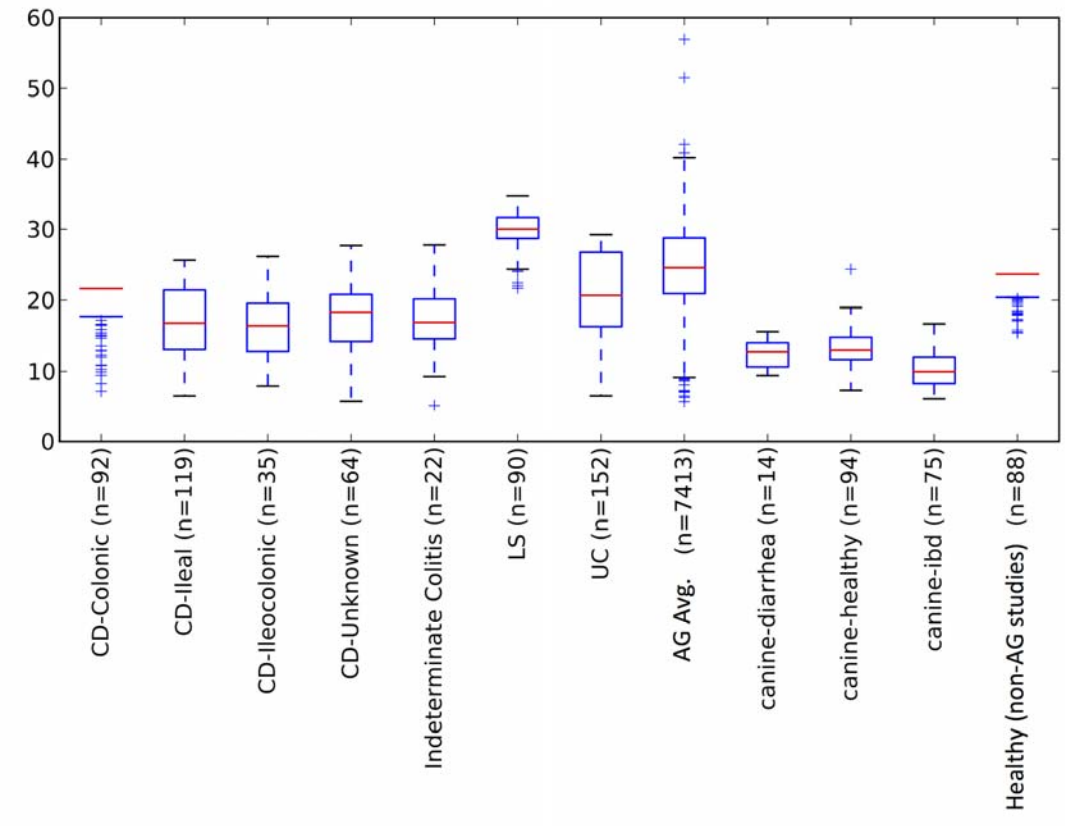
Figure 5. The mean relative abundance of bacterial phyla present in each disease subtype is illustrated.

As mentioned above, a phylum level analysis is shallow in the sense that it looks at the broadest taxonomic classification available. To note finer differences between datasets, looking at deeper taxonomic levels-such as genus-can enable researchers to make observations not possible at deeper taxonomic levels. For example, we do not observe any notable differences between the disease sub-types in this meta-analysis at the phylum level, but looking at the top ten genera (plural of genus) present in each disease sub-type (**Table 1**) reveals some differences; for example, *Haemophilus* and *Veillonella* are among the top ten genera present in colonic CD, but not in ileal, ileocolonic, or uncharacterized CD. Additionally, *Parabacteroides* and *Morganella* were among the top ten genera in the Smarr cohort but not in any other cohort studied here-further highlighting the separation of this individual from other IBD patients.

CD-ileal	**CD-ileocolonic**	**CD-colonic**	**CD-Unk**	**UC**	**Ind. colitis**	**LS**	**Healthy (non-AG groups)**
Bacteroides (16.2%)	Bacteroides (19.4%)	Bacteroides (22.5%)	Bacteroides (12.9%)	Bacteroides (17.4%)	Bacteroides (18.4%)	Bacteroides (21.3%)	Bacteroides (29.3%)
Lachnospiraceae (9.5%)	Lachnospiraceae (11.3%)	Lachnospiraceae (9.8%)	Faecalibacterium (7.6%)	Faecalibacterium (7.9%)	Faecalibacterium (9.8%)	Ruminococcaceae (11.4%)	Lachnospiraceae (12.7%)
Blautia (7.5%)	Faecalibacterium (5.4%)	Blautia (6.5%)	Lachnospiraceae (7.2%)	Lachnospiraceae (7.5%)	Blautia (9.3%)	Lachnospiraceae (8.9%)	Prevotella (6.8%)
Ruminococcus (5.6%	Blautia (5.3%)	Streptococcus (5.4%)	Blautia (7.2%)	Blautia (6.7%)	Lachnospiraceae (6.9%)	Faecalibacterium (8.8%)	Ruminococcaceae (6.5%)
Streptococcus (4.4%)	Ruminococcaceae (4.4%)	Faecalibacterium (5.2%)	Ruminococcaceae (6.3%)	Ruminococcaceae (6.5%)	[Ruminococcus] (Lacho) (6.0%)	Unk. Genus, order Clostridiales (5.7%)	Faecalibacterium (5.8%)
Faecalibacterium (3.7%)	Prevotella (4.4%)	Ruminococcaceae (4.3%)	[Ruminococcus] (Lachno) (5.1%)	Bifidobacterium (3.6%)	Ruminococcaceae (5.4%)	Rikenellaceae (4.4%)	Unk. Genus, order Clostridiales (2.5%)
Ruminococcaceae (3.5%)	Akkermansia (4.0%)	Ruminococcus (3.8%)	Bifidobacterium (4.4%)	Coprococcus (3.2%)	Ruminococcus (Rumino) (4.3%)	Akkermansia (4.3%)	Ruminococcus (Rumino) (2.4%)
Dialister (3.4%)	Dialister (3.9%)	Haemophilus (2.7%)	Streptococcus (4.3%)	Unk. Genus, order Clostridiales (3.1%)	Bifidobacterium (4.3%)	Parabacteroides (4.0%)	Blautia (2.3%)
Clostridium (2.9%)	Ruminococcus (3.6%)	Veillonella (2.6%)	Unk. Genus, order Clostridiales (3.6%)	Ruminococcus (Rumino) (3.1%)	Coprococcus (3.7%)	Morganella (3.7%)	[Ruminococcus] (Lachno) (2.3%)
Bifidobacterium (2.7%)	Bifidobacterium (3.3%)	Coprococcus (2.6%)	Coprococcus (3.2%)	[Ruminococcus] (Lachno) (2.8%)	Unk. Genus, order Clostridiales (3.3%)	Prevotella (3.3%)	Coprococcus (1.9%)

Table 1. The top ten genera present in each disease subtype, along with the mean relative abundance of each genus, are listed.

Alpha diversity is another way to assess microbiome data. In simple terms, alpha diversity is an assessment of "who" is in a sample-i.e. which individual bacterial taxa are present in a sample (for example, how many different species are in a sample). We use a metric called Phylogenetic Diversity to assess the alpha diversity of a sample as it typically better preserves feature diversity than does simply calculating the number of species present. The best way to understand how this metric works is to think about the following: two samples, each with 5 taxa, might be considered equally diverse (the number of taxa in each is the same); But, if one of those samples is comprised only of *Bacteroides* species and the other contains a mix of *Bacteroides, Prevotella*, and *Ruminococcus* species, the second sample is more phylogenetically diverse than the one comprised only of *Bacteroides*-and therefore will receive a higher diversity score. In most cases, increased microbial community diversity is an indicator of a healthy state, and diseased states are often associated with decreased diversity. Several disease populations have been found to have lower alpha diversity than controls, including people with obesity, Type I and Type II diabetes, and several forms of IBD. People who eat a lot of fermented foods, diverse plant-based diets, and who live hunter-gatherer lifestyles have especially high alpha diversity. Further research is required to test whether change in alpha diversity is a cause or consequence of disease, and whether increasing an individual's alpha diversity has benefits per se. In the meta-analysis presented here, we determined alpha diversity for samples grouped by disease subtype, calculating the averages for each group. Below, a box and whisker plot illustrates differential alpha diversity between different disease subtypes.



As expected, diversity is decreased in CD and UC patients compared to healthy individuals, and the American Gut cohort. Diversity is slightly lower in CD than in UC; however, no differences are significant. This is likely due to low samples sizes in some of the groups. Interestingly, the Smarr dataset has the highest diversity (even higher than healthy subjects), supporting the PCoA plots, which reveal that these samples do not cluster with the other IBD cohorts.

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

These data were presented at Digestive Disease Week® 2016 at the session titled "Active Learning Session on the Gut Microbiome: Fundamentals of Theory and Practice." Visit www.ddw.org for more information.

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

Bibliography

[1] "American Gut," 2016. [Online]. Available: http://americangut.org.

[2] "Human Microbiome Project," 2016. [Online]. Available: https://commonfund.nih.gov/hmp/index.

[3] "Earth Microbiome Project," 2011. [Online]. Available: http://earthmicrobiome.org.

[4] D. Gevers, S. Kugathasan, L. A. Denson, Y. Vazquez-Baeza, W. Van Treuren, B. Ren, E. Schwager, D. Knights, S. J. Song, M. Yassour, X. C. Morgan, A. D. Kostic, C. Luo, A. Gonzalez, D. McDonald, Y. Haberman, T. Walters, S. Baker, J. Rosh, M. Stephens, M. Heyman, J. Markowitz, R. Baldassano, A. Griffiths, F. Sylvester, D. Mack, S. Kim, W. Crandall, J. Hyams, C. Huttenhower, R. Knight and R. J. Xavier, "The treatment-naive microbiome in new-onset Crohn's disease," *Cell Host Microbe,* vol. 15, no. 3, pp. 382-392, 2014.