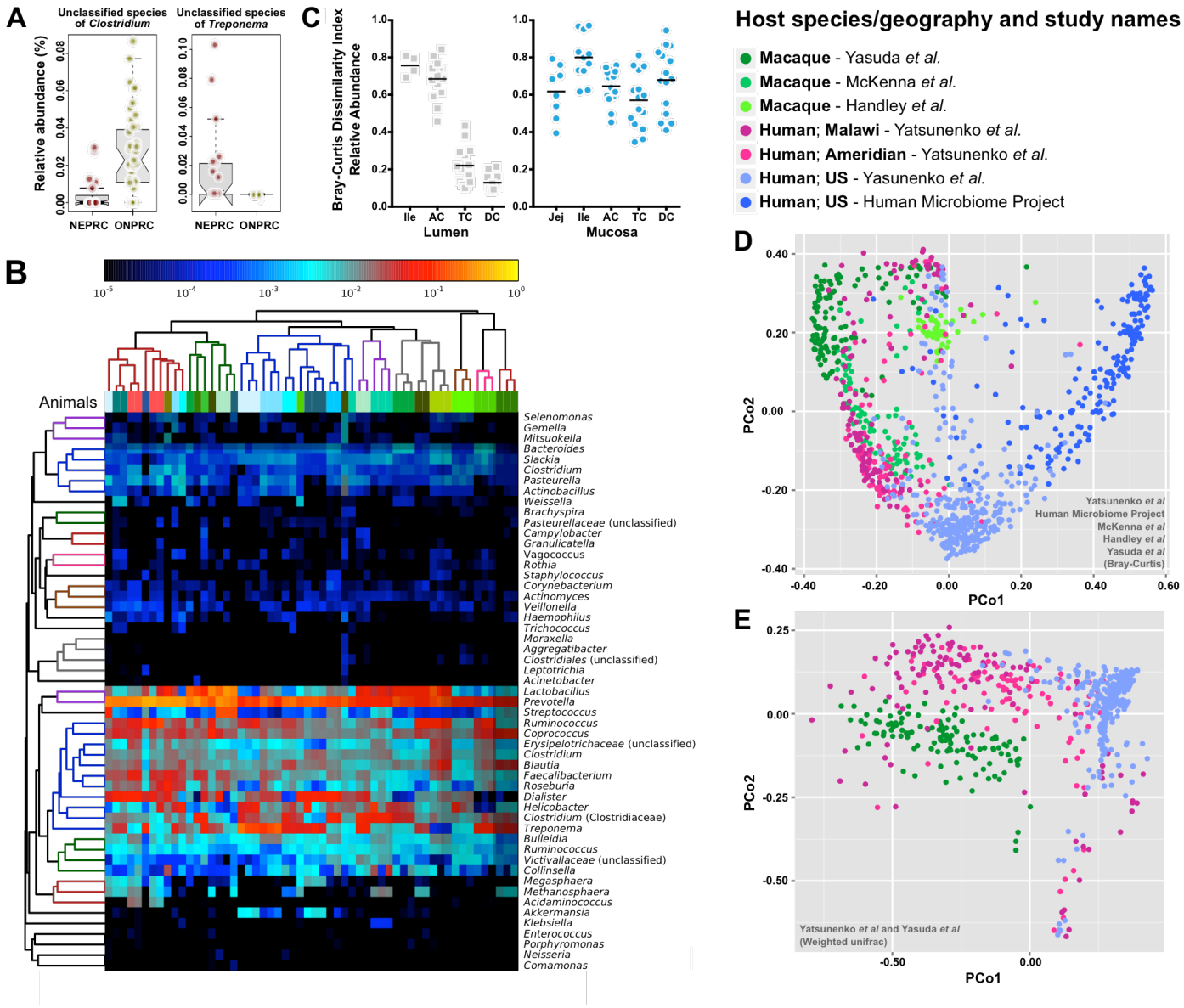

Figure S1, related to Figure 1: Influences on gut microbial composition and relating macaque and human microbiota A) A multivariate analysis identified twenty-three taxa that were differentially abundant between the source primate centers from which our cohort originated. Two examples are shown here: an unclassified species of *Treponema,* and an unclassified species of *Clostridium.* The complete list of differentially expressed taxa is available in **Table S1**. B) The Bray-Curtis distance between each sample and the stool sample of the same macaque is plotted for lumenal (left, gray) and mucosal (right, light blue) samples. Samples are stratified by intestinal region. C) To address the influence of host on microbial diversity, all colonic lumen and stool samples are hierarchically clustered based on Bray-Curtis dissimilarity index. Top bar indicates individual animal. D and E) The similarity of microbial communities described in this study, two other macaque studies (McKenna et al., 2008; Handley et al., 2012) and two human studies (Human Microbiome Project, 2012; Yatsunenko et al., 2012) was assessed by calculating the Bray-Curtis dissimilarity and weighted Unifrac distances, then performing principal coordinate analysis. D) Community distance was measured by Bray-Curtis dissimilarity. This plot includes all five studies. E) Community distance was measured by weighted Unifrac distance. This plot only includes the Yatsunenko et al and Yasuda et al datasets.

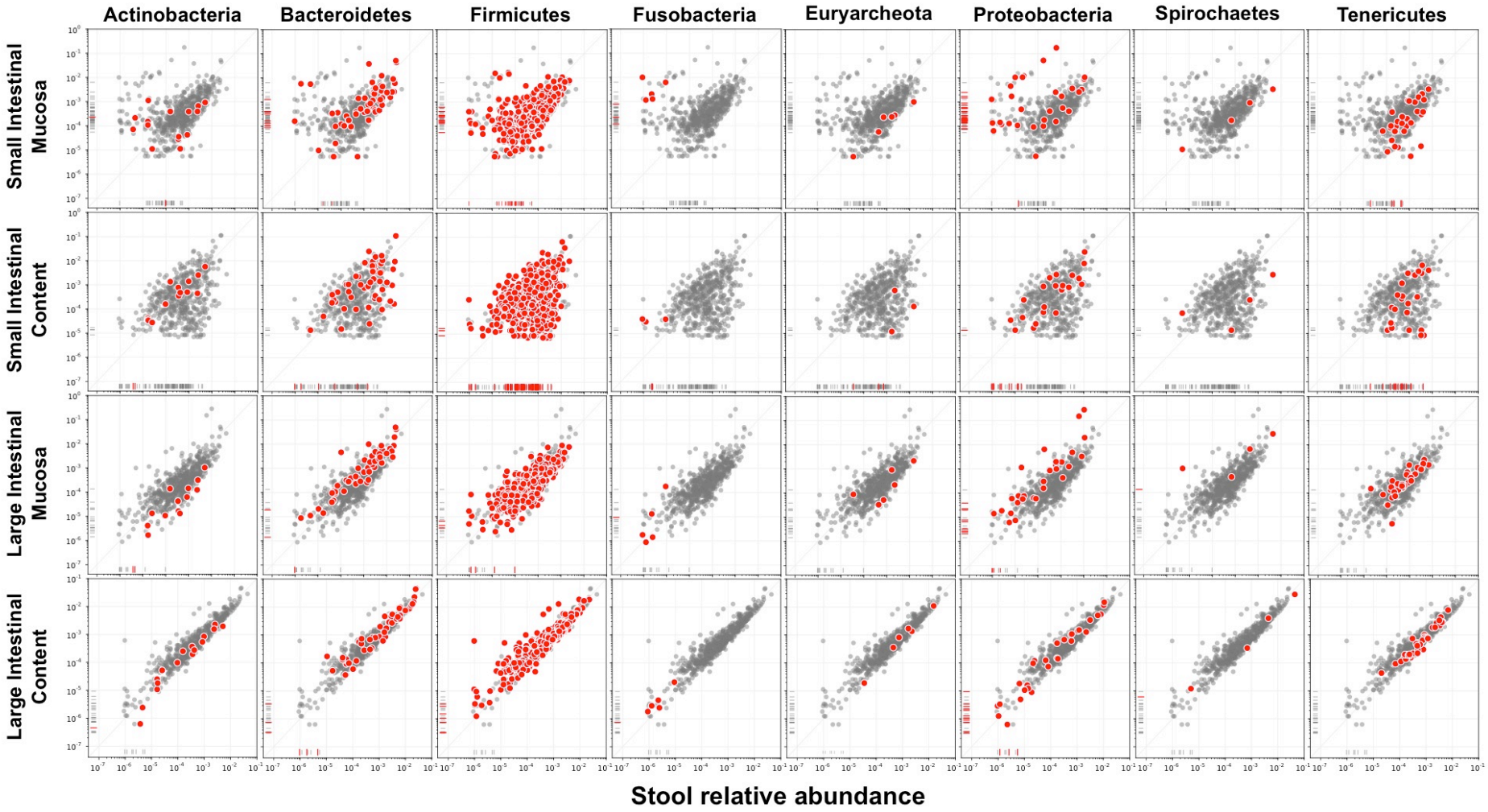

Figure S2, related to Figure 2: A phylum-level view of mucosal taxa underrepresented in stool Each dot corresponds to the average relative abundance of an OTU across 15 animals in each intestinal region (SI mucosa and content, LI mucosa and content). Clades of interest are highlighted in red. Marks on the x-axis (vertical lines) or y-axis (horizontal lines) margins represent OTUs with zero measured abundance at one site but non-zero abundance at the other.

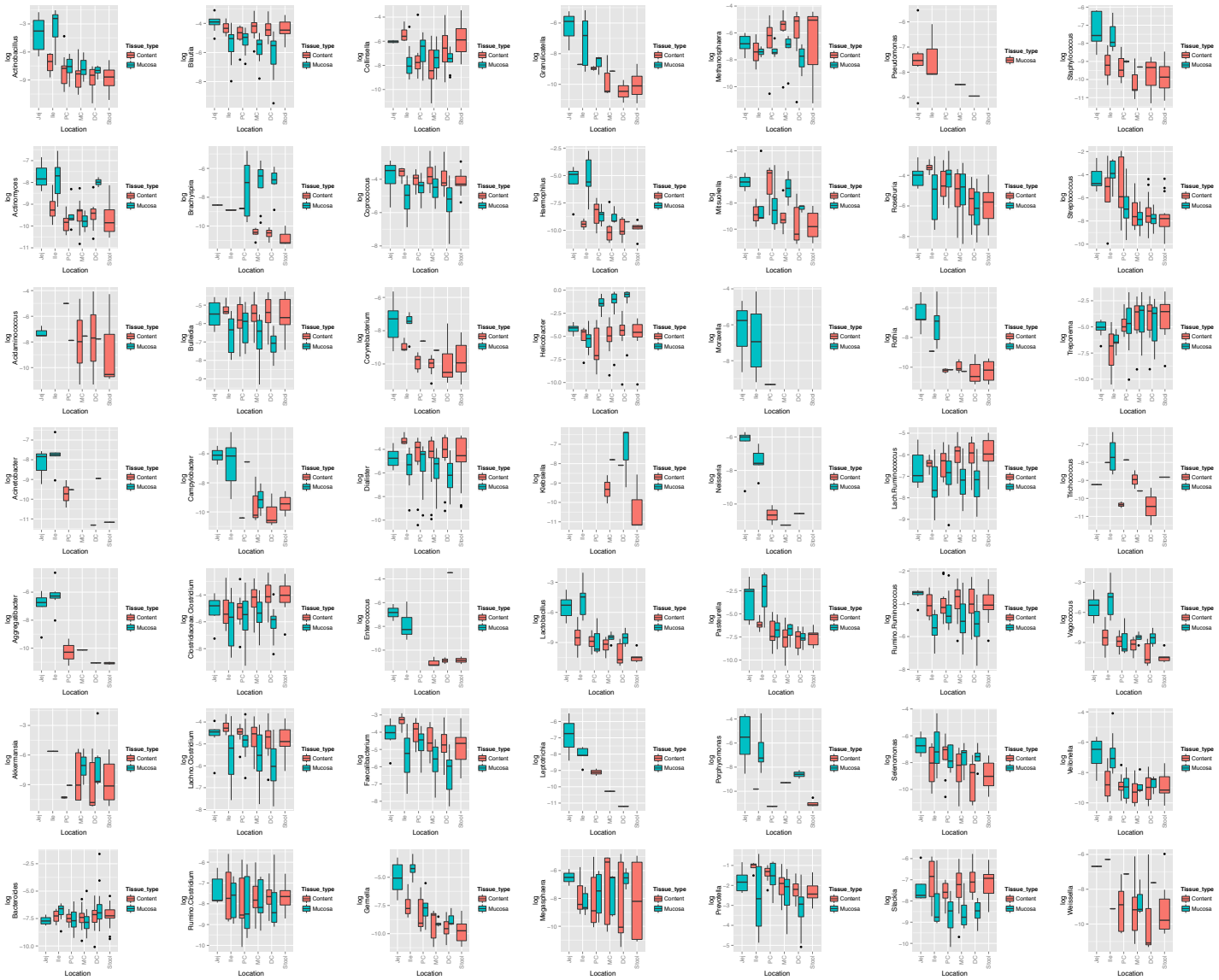

Figure S3, related to Figure 4: Biogeographic distribution of taxa throughout the gut For each genus, distribution of its mucosal, lumenal, and stool abundance is shown across the population of 15 macaques, stratified by geography (jejunum, ileum, proximal colon, mid colon, distal colon, stool) and mucosa / lumena. All mucosal samples are shown in blue (except *Pseudomonas*); all lumenal/stool samples are shown in red, and all y axes are log relative abundance. Jejunum and stool samples appear twice as wide as the others because there is not a corresponding paired sample – jejunum is mucosal-only, and stool is stool-only.

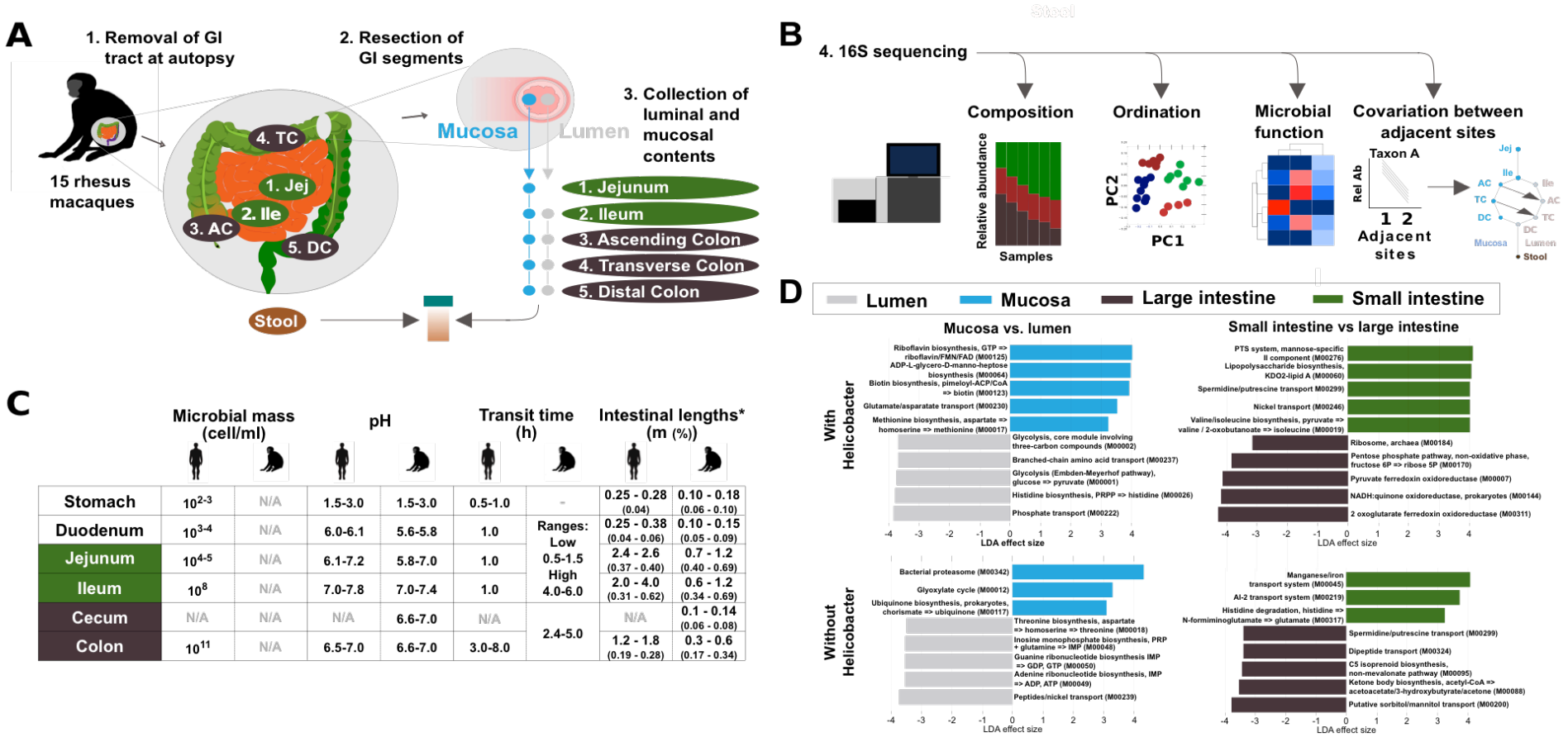

Figure S4, related to Experimental Procedures. Study design and survey of primate gut microbial biogeography and microbial functional potentials with and without *Helicobacter* A) Paired intestinal mucosal and lumenal contents were collected from the ileum, ascending, transverse, and descending colon of 15 clinically-healthy rhesus macaques, in addition to stool and a sample of jejunal mucosa. The microbiome of the samples was profiled by sequencing the V4 region of the 16S rRNA gene. B) After sequencing, community structure, function, and covariation with biogeography were characterized by ordination (Caporaso et al., 2010), univariate (Segata et al., 2011a) and multivariate (Morgan et al., 2012) association testing, metagenomic inference (Langille et al., 2013), and logistic regression. C) Comparison of the gastrointestinal tracts of humans and rhesus macaques. In contrast to macaques, humans lack a prominent cecum. The total length of the GI tract is 6-7 m for an adult human and 1.5-2m for an adult rhesus macaque. Comparison of intestinal microbial mass (Solnick et al., 2006; Walter and Ley, 2011), pH (Mercier et al., 2007; Walter and Ley, 2011) and transit time (Dubois et al., 1977; Mercier et al., 2007). Percent of intestinal length is normalized to an intestinal length of 6.5 m for humans and 1.75 m for a macaque, for comparison purposes. D) PICRUSt (Langille et al., 2013) was used to infer community function, and LEFSe (Segata et al., 2011b) was used to determine which functions were most differential between the mucosa and lumen and LI and SI. Due to the high abundance of *Helicobacter*, this analysis was repeated with *Helicobacter* removed. The ten largest LDA effects are shown here. The top and bottom two panels are derived from 16S data including *Helicobacte*r, and excluding *Helicobacter* OTUs, respectively.

Table captions

Table S1, associated with Figure S1 and Experimental Procedures Bacterial taxa and functions significantly enriched by multivariate analysis in either the mucosa or lumen, a location, or a primate center of origin. Functional analysis was performed with and without Helicobacter.

Table S2, associated with Figure 3 Bacterial OTUs identified in 4 major regions of the intestine but not identified in stool

Supplemental Experimental Procedures

Animals and sample collection

Cohort Metadata	
n	15
Gender	Female
Average age ± s.d. (Range)	18 ± 3.5 (13, 22)
Average body weight ± s.d. (Range)	10kg ± 1.6 (6.9, 12.5)
Diet[1]	Adult monkey chow
Health status	Clinically healthy

Sampling locations

	Mucosa (n)	Lumen (n)
Jejunum (Jej)	8	-
Ileum (Ile)	11	4
Ascending colon (AC)	14	15
Transverse colon (TC)	15	14
Descending colon (DC)	15	15
Stool	-	15

All animals that were housed at the NEPRC in accordance with all applicable regulations and in a facility accredited by the Association for Assessment and Accreditation of Laboratory Animal Care International. Animals were maintained under an experimental protocol approved by Harvard Medical School's Standing Committee on Animals. Prior to sample collection, animals were housed and fed individually.

Intestinal lumen (ileum, ascending, transverse, and distal colon), mucosal

scrapings (jejunum, ileum, ascending, transverse, descending colon), and stool samples were collected during autopsy from 15 clinically-healthy female rhesus macaques, ranging from 12 to 22-years old (See Table Cohort Metadata).

The entire intestinal tract was first removed from the body. Next, a 15-cm section from each biogeographical location was cross-sectionally transected, and then longitudinally transected on the anti-mesenteric side of the intestine to open the intestinal lumen (**Fig. S4**). Lumenal samples were collected by advancing the lumenal contents into a cryotube (Nunc CryTubes, Sigma-Aldrich, St. Louis, MO) using a sterile spatula. Intestinal contents were removed from the lumen and rinsed with sterile saline to remove any visible contents without disturbing the intestinal mucosa. It was not possible to collect jejunal lumenal contents due to fasting of the animals prior to euthanasia. Intestinal mucosal samples were then collected by gently scraping the mucosal surface with a sterile glass slide (to avoid penetrating the basement membrane) and scraped samples were advanced to a cryotube. All intestinal samples were snap frozen in liquid nitrogen and stored at -80°C for further analysis. All histopathology of the intestinal tissues and major organs was normal.

16S rRNA sequencing and profiling

DNA from stool, mucosal, and lumenal samples was extracted using the MP BIO FASTDNA™ SPIN Kit for Soil (MP Bio, Santa Ana, CA) according to manufacturer's instructions. The amplification and sequencing of the V4 region by Illumina MiSeq were performed as described previously (Yatsunenko et al., 2012). In brief, genomic DNA was subjected to 16S amplifications using primers designed incorporating the Illumina adapters and a sample barcode sequence, allowing directional sequencing covering variable region V4 (Primers: 515F [GTGCCAGCMGCCGCGGTAA] and 806R [GGACTACHVGGGTWTCTAAT]). PCR

mixtures contained 10 µl of diluted template (1:50), 10 µl of HotMasterMix with the HotMaster Taq DNA Polymerase (5 Prime), and 5 µl of primer mix (2 µM of each primer). The cycling conditions consisted of an initial denaturation of 94°C for 3 min, followed by 30 cycles of denaturation at 94°C for 45 sec, annealing at 50 °C for 60 sec, extension at 72°C for 5 min, and a final extension at 72°C for 10 min.

Amplicons were quantified on the Caliper LabChipGX (PerkinElmer, Waltham, MA), pooled in equimolar concentrations, size selected (375-425 bp) on the Pippin Prep (Sage Sciences, Beverly, MA) to reduce non-specific amplification products from host DNA, and a final library size and quantification was performed on an Agilent Bioanalyzer 2100 DNA 1000 chips (Agilent Technologies, Santa Clara, CA). Sequencing was performed on the Illumina MiSeq platform (version 2) according to the manufacturer's specifications with addition of 5% PhiX, and yielded paired-end reads of 175 bp in length in each direction.

16S sequence bioinformatic processing

Overlapping paired-end reads were stitched together (approximately 97 bp overlap), size-selected to reduce non-specific amplification products from host DNA (225 - 275 bp), and further processed in a data curation pipeline implemented for PICRUSt (Langille et al., 2013) in QIIME 1.6.0 as pick_closed_reference_otus.py (Caporaso et al., 2010). In brief, this pipeline will (i) pick OTUs using a reference-based method and then (ii) constructs an OTU table. Taxonomy is assigned using the Greengenes (18 May 2012 version) predefined taxonomy map of reference sequence OTUs to taxonomy (McDonald et al., 2012). The resulting OTU tables are checked for mislabeling and contamination (Knights et al., 2011).

A mean sequence depth of 29,914/sample was obtained; samples with fewer than 3,000 filtered sequences and those Operational Taxonomic Units (OTUs) with less than 15 reads were excluded from downstream analysis. Further microbial community analysis such as beta diversity was calculated with QIIME 1.6.0 (Caporaso et al., 2010). To test for statistically significant association between the microbiota and metadata including biogeographical locations, we used LEfSe (Segata et al., 2011) for univariate and MaAsLin (Multivariate Associations by Linear models) (Morgan et al., 2012) for multivariate analyses (**Table S1**). We used LEfSe to identify features (microbial taxa) that separate two classes (mucosa vs. lumen or small vs. large intestine) and quantify effect sizes (i.e. biological magnitude) of the association. We used MaAsLin to build a multivariate linear model combining fixed and random effects to identify associations between microbial communities with covariates including sample type (mucosa vs. lumen), locations (jejunum, ileum, ascending, transverse, and distal colon, and stool), age, body weight, and primate center origin). We controlled for individuals. For MaAsLin data, we used Benjamini-Hochberg false discovery rate corrections to accept no more than a 20% FDR.

In order to predict microbial functions from the microbial data, we used PICRUSt (Langille et al., 2013). This algorithm estimates the functional potential of microbial communities given a marker gene survey and the set of currently-sequenced reference genomes with an accuracy of 80-90% on human gut communities. Although predicted metagenomes derived from PICRUSt provide informative functionalities of the microbial community, they are often specific (e.g. glycerol-3-phosphate dehydrogenase (NAD+)). We thus used HUMAnN (Abubucker et al., 2012) to identify KEGG modules (version 56) based on the metagenome predicted from the 16S sequencing data using PICRUSt. KEGG module is a collection of manually-defined functional units and can be used to interpret biological functions of

metagenomic data. The result of the univariate (LEfSe) and multivariate (MaAsLin) analyses are included in **Fig. S4D** and **Table S1**.

To assess the similarity of our data to previously-published macaque and human studies, microbiota data, either taxonomic or raw sequencing data were obtained from publically available sources (Handley et al., - RG-RAST: http://metagenomics.anl.gov/?page=MetagenomeSelect; Human Microbiome Project (HMP) – http://www.hmpdacc.org/reference_genomes/reference_genomes.php; Yatsunenko - https://gordonlab.wustl.edu/SuppData.html) or directly from the investigator (McKenna et al., 2008). Taxonomic tables were summarized to genus-level clades and merged. All studies except for Yatsunenko et al and the current study used different PCR amplification methods, sequencing platforms, and variable regions of the 16S rRNA gene (see below). The Bray-Curtis distance was used to assess the similarity between all five communities (Figure S1D). Since Yatsunenko et al. and the current study used the same methods to amplify, sequence, and assign taxonomy, the weighted Unifrac distance, which measures the phylogenetic relatedness as well as the counts of each taxa, was used to assess similarity between the Yatsunenko dataset and the current study (Figure S1E).

Study Name	**Host Species**	**Sequence**	**Method**	**Regions**	**Sample Type**
Human Microbiome Project	Human - US	454	16S	V1-V3, V3-V5	Stool
Yatsunenko *et al*	Human - US	Illumina	16S	V4	Stool
Yatsunenko *et al*	Human - Ameridian	Illumina	16S	V4	Stool
Yatsunenko *et al*	Human - Malawi	Illumina	16S	V4	Stool
McKenna *et al*	Macaque	454	16S	V1-3	Stool + Biogeography
Handley *et al*	Macaque	454	Shotgun metagenome		Stool

Yasuda *et al*	Macaque	Illumina	16S	V4	Stool + Biogeography

Identification of microbial taxa enrichment sites and predictability by logistic regression

For each OTU,

$$\log\frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k S_1 + \beta_{k+1} S_2 + \cdots$$

$$Y \sim Binomial(\, N, p \,)$$

Where

p, proportion of this OTU at this location;

X_i, indicator variable for location;

S_i, indicator variable for subject;

Y, reads corresponding to this OTU at this location for this subject;

N, reads for all OTUs at this location for this subject

The *Circular layout* option included in Cytoscape (Cline et al., 2007) (Cytoscape version 3.0.1.) was used to visualize the predictability of microbial taxa between adjacent biogeographical sites for each taxa. The direction of β (positive, negative, and none-significant) was used as the type of interaction, and attributes included relative abundance of each taxa at each location, and magnitude of β derived above. Although in some cases when abundances of distal sites are higher than proximal sites (i.e. abundances in stool are higher than distal colon lumen), in those cases, the negative βs suggested that this bacterial taxa may go from stool to distal colon, the fact that this is unlikely in reality considering the natural flow of intestinal contents. Therefore, when the direction of β (either positive, or negative) opposed the actual

physiological flow (we assumed the actual physiological flow to be always proximal to distal amongst mucosa and lumen and interchangeable between mucosa and lumen), the errors were substituted with lines and combined with the non-significant group, which was also noted as a line.

Supplemental References

Abubucker et al., 2012 Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Comput Biol *8*, e1002358.

Caporaso et al., 2010 QIIME allows analysis of high-throughput community sequencing data. Nat Methods *7*, 335-336.

Cline et al., 2007 Integration of biological networks and gene expression data using Cytoscape. Nature protocols *2*, 2366-2382.

Dubois et al., 1977 Gastric emptying and secretion in the rhesus monkey. The American journal of physiology *232*, E186-192.

Knights et al., 2011 Supervised classification of microbiota mitigates mislabeling errors. The ISME journal *5*, 570-573.

McDonald et al., 2012 The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. GigaScience *1*, 7.

Solnick et al., 2006 Acquisition of Helicobacter pylori infection in rhesus macaques is most consistent with oral-oral transmission. Journal of clinical microbiology *44*, 3799-3803.