Gut Microbiota Diversity across Ethnicities in the United States

Authors:

Andrew W. Brooks[1,2], Sambhawa Priya[3,4,5], Ran Blekhman[3,4], Seth R. Bordenstein[1,2,6,7*]

Affiliations:

[1] Vanderbilt Genetics Institute, Vanderbilt University, Nashville, TN, USA.

[2] Department of Biological Sciences, Vanderbilt University, Nashville, TN, USA

[3]Department of Genetics, Cell Biology, and Development, University of Minnesota, Minneapolis, MN, USA

[4]Department of Ecology, Evolution, and Behavior, University of Minnesota, Minneapolis, MN, USA

[5] Bioinformatics and Computational Biology Program, University of Minnesota, Minneapolis, MN, USA

[6] Department of Pathology, Microbiology, and Immunology, Vanderbilt University, Nashville, TN, USA.

[7]Vanderbilt Institute for Infection, Immunology and Inflammation, Vanderbilt University, Nashville, TN, USA.

[*] Corresponding author. E-mail: s.bordenstein@vanderbilt.edu (S.R.B)

18 **Abbreviations:**

19 AGP – American Gut Project

20 ANOSIM – Analysis of Similarity

21 AUC – Area Under the Curve

22 BMI – Body Mass Index

23 F_{ST} – Fixation Index

24 GWAS - Genome-Wide Association Studies

25 HMP – Human Microbiome Project

26 MAF - Minor Allele Frequency

27 OTU – Operational Taxonomic Unit

28 PERMANOVA - Permutational Multivariate Analysis of Variance

29 RF – Random Forest

30 ROC – Receiver Operating Characteristic

31 SMOTE – Synthetic Minority Over-sampling Technique

Abstract:

Composed of hundreds of microbial species, the composition of the human gut microbiota can vary with chronic diseases underlying health disparities that disproportionally affect ethnic minorities. However, the influence of ethnicity on the gut microbiota remains largely unexplored and lacks reproducible generalizations across studies. By distilling associations between ethnicity and differences in two United States based 16S gut microbiota datasets including 1,673 individuals, we report 12 microbial genera and families that reproducibly vary by ethnicity. Interestingly, a majority of these microbial taxa, including the most heritable bacterial family, Christensenellaceae, overlap with genetically-associated taxa and form co-occurring clusters linked by similar fermentative and methanogenic metabolic processes. These results demonstrate recurrent associations between specific taxa in the gut microbiota and ethnicity, providing hypotheses for examining specific members of the gut microbiota as mediators of health disparities.

45 **Introduction:**

46 The human gut microbiota at fine resolution varies extensively between individuals
47 (1-3), and this variability frequently associates with diet (4-7), age (6, 8, 9), sex (6, 9, 10),
48 body mass index (BMI) (1, 6), and diseases presenting as health disparities (11-14). The
49 overlapping risk factors and burden of many chronic diseases disproportionally affect ethnic
50 minorities in the United States, yet the underlying biological mechanisms mediating these
51 substantial disparities largely remain unexplained. Recent evidence is consistent with the
52 hypothesis that ethnicity associates with variation in microbial abundance, specifically in the
53 oral cavity, gut, and vagina (15-17). To varying degrees, ethnicity can capture many facets of
54 biological variation including social, economic and cultural variation, as well as aspects of
55 human genetic variation and biogeographical ancestry. Ethnicity also serves as a proxy to
56 characterize health disparity incidence in the United States, and while factors such as genetic
57 admixture create ambiguity of modern ethnic identity, self-declared ethnicity has proven a
58 useful proxy for genetic and socioeconomic variation in population scale analyses, including
59 in the Human Microbiome Project (18-20). Microbiota differences have been documented
60 across populations that differ in ethnicity as well as in geography, lifestyle, and sociocultural
61 structure; however, these global examinations cannot disconnect factors such as
62 intercontinental divides and hunter-gatherer versus western lifestyles from ethnically
63 structured differences (21-23). Despite the importance of understanding the
64 interconnections between ethnicity, microbiota, and health disparities, there are no
65 reproducible findings about the influence of ethnicity on differences in the gut microbiota
66 and specific microbial taxa in diverse United States populations, even for healthy individuals
67 (6).

68 Here, we comprehensively examine connections between self-declared ethnicity and
69 gut microbiota differences across more than a thousand individuals sampled by the
70 American Gut Project (AGP, N=1375) (24) and the Human Microbiome Project (HMP, N=298)
71 (6). Previous studies demonstrated that human genetic diversity in the HMP associates with
72 differences in microbiota composition(25), and genetic population structure within the HMP
73 generally delineates self-declared ethnicity (20). Ethnicity was not found to have a
74 significant association with microbiota composition in a Middle Eastern population,
75 however factors such as lifestyle and environment that influence microbiota variation across
76 participants was homogenous compared to the ethnic, sociocultural, economic, and dietary
77 diversity found within the United States (26). While ethnic diversity is generally
78 underrepresented in current microbiota studies, evidence supporting an ethnic influence on
79 microbiota composition among first generation immigrants has been recently demonstrated
80 in a Dutch population (27). The goal of this examination is to evaluate, for the first time, if
81 there are reproducible differences in gut microbiota across ethnicities within an overlapping
82 United States population, as ethnicity is one of the key defining factors for health disparity
83 incidence in the United States. Lifestyle, dietary, and genetic factors all vary to different
84 degrees across ethnic groups in the United States, and it will require more even sampling of
85 ethnic diversity and stricter phenotyping of study populations to disentangle which factors
86 underlie ethnic microbiota variation in the AGP and HMP.

87 **Results:**

88 ***Microbiota are subtly demarcated by ethnicity***

89 We first evaluate gut microbiota distinguishability between AGP ethnicities (**Fig 1A**,
90 family taxonomic level, Asians-Pacific Islanders (N=88), Caucasians (N=1237), Hispanics
91 (N=37), and African Americans (N=13)), sexes (female (N=657), male (N=718)), age groups
92 (years grouped by decade), and categorical BMI (underweight (N=70), normal (N=873),
93 overweight (N=318), and obese (N=114)) (Demographic details in **S1A Table**). Age, sex, and
94 BMI were selected as covariates because they are consistent across the AGP and HMP
95 datasets. Additionally, 31 other categorical factors measuring diet, environment, and
96 geography were compared for pairwise differences between two ethnicities using
97 proportions tests, and very few (10 / 894) tests significantly varied (**S1 Table** additional
98 sheets). Interindividual gut microbiota heterogeneity clearly dominates; however, Analyses
99 of Similarity (ANOSIM) reveal subtle but significant degrees of total microbiota
100 distinguishability for ethnicity, BMI, and sex, but not for age (**Fig 1B**, Ethnicity; **Fig 1C**, BMI;
101 **Fig 1D**, Sex; **Fig 1E**, Age) (28). Recognizing that subtle microbiota distinguishability between
102 ethnicities may be spurious, we independently replicate the ANOSIM results from HMP
103 African Americans (N=10), Asians (N=34), Caucasians (N=211) and Hispanics (N=43) (**S2A**
104 **Table**, R=0.065, p=0.044). We again observe no significant distinguishability for BMI, sex,
105 and age in the HMP. Higher rarefaction depths increase microbiota distinguishability in the
106 AGP across various beta diversity metrics and categorical factors (**S2B Table**), and
107 significance increases when individuals from overrepresented ethnicities are subsampled
108 from the average beta diversity distance matrix (**S2C Table**). Supporting the ANOSIM results,
109 Permutational Multivariate Analysis of Variance (PERMANOVA) models with four different

110 beta diversity metrics showed that while all factors had subtle but significant associations
111 with microbiota variation when combined in a single model, effect sizes were highest for
112 ethnicity in 7 out of 8 comparisons across beta diversity metrics and rarefaction depths in
113 the AGP and HMP (**S2D Table**). We additionally test microbiota distinguishability by
114 measuring the correlation between beta diversity and ethnicity, BMI, sex, and age with an
115 adapted BioEnv test (**S2E Table**) (29). Similar degrees of microbiota structuring occur when
116 all factors are incorporated (Spearman Rho=0.055, p-values: Ethnicity=0.057, BMI<0.001,
117 Sex<0.001, Age=0.564). Firmicutes and Bacteroidetes dominated the relative phylum
118 abundance, with each representing between 35% and 54% of the total microbiota across
119 ethnicities (**S1 Fig**).

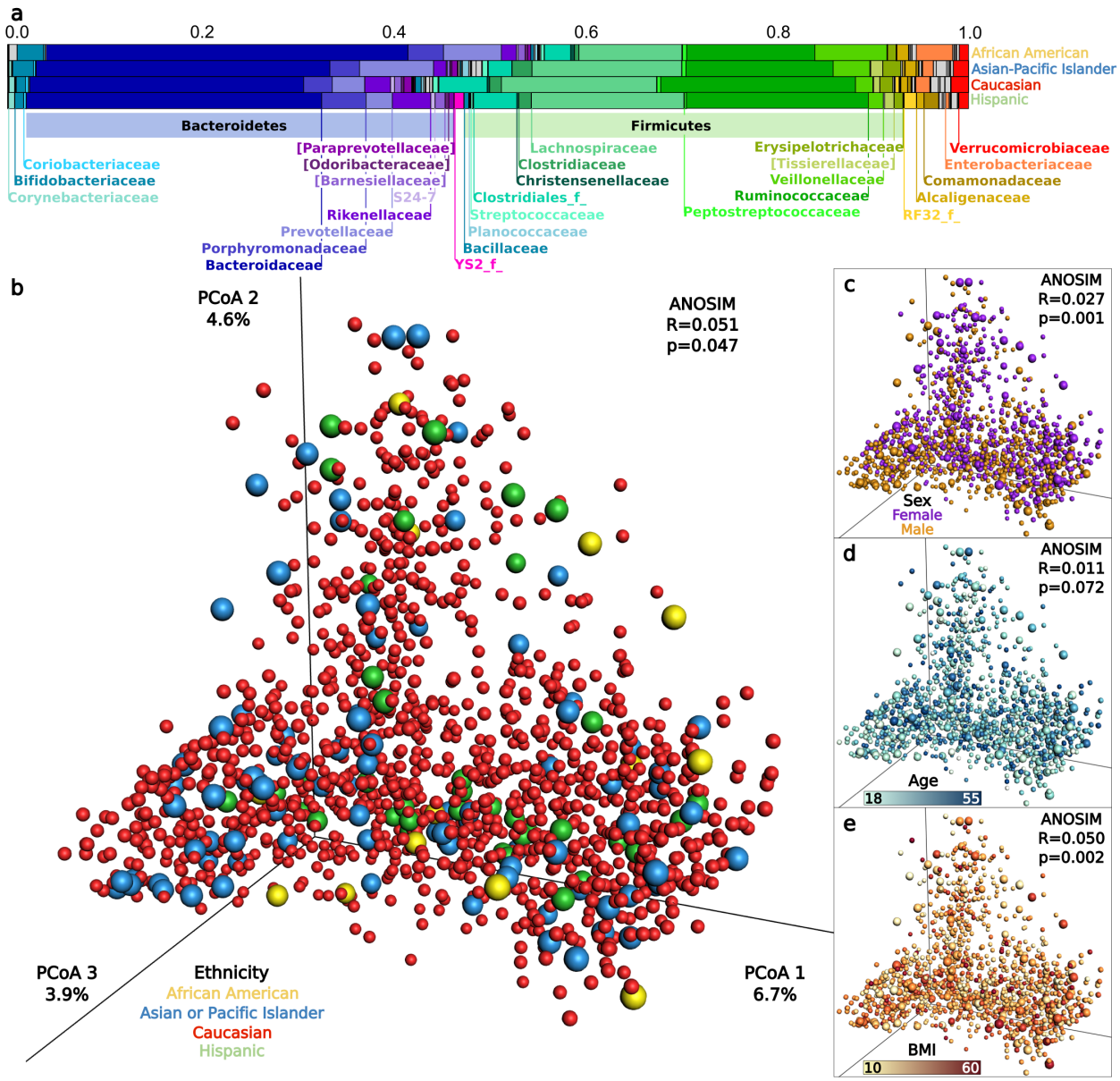


Fig 1. Gut microbiota composition and distinguishability by ethnicity, sex, age and BMI. (A) The average relative abundance of dominant microbial families for each ethnicity. (B-E) Principle coordinates analysis plots of microbiota Bray-Curtis beta diversity and ANOSIM distinguishability for: (B) Ethnicity, (C) Sex, (D) Age, (E) BMI. In B-E, each point represents the microbiota of a single sample, and colors reflect metadata for that sample.

126 Caucasian points are reduced in size to allow clearer visualization, and p-values are not
127 corrected across factors which have different underlying population distributions.
128
129 We next test for ethnicity signatures in the gut microbiota by analyzing alpha and beta
130 diversity, abundance and ubiquity distributions, distinguishability, and classification
131 accuracy (30). Shannon's Alpha Diversity Index (31), which weights both microbial
132 community richness (Observed OTUs) and evenness (Equitability), significantly varies
133 across ethnicities in the AGP dataset (Kruskal Wallis, p=2.8e-8) with the following ranks:
134 Hispanics > Caucasians > Asian-Pacific Islanders > African Americans (**Fig 2A**). In the HMP,
135 there is a significantly lower Shannon diversity for Asian-Pacific Islanders relative to
136 Caucasians and a trend of lower Shannon diversity for Asian-Pacific Islanders relative to
137 Hispanics; African Americans change position in diversity relative to other ethnicities,
138 potentially as a result of undersampling bias. Five alpha diversity metrics, two rarefaction
139 depths, and separate analyses of Observed OTUs and Equitability generally confirm the
140 results (**S3A Table**).

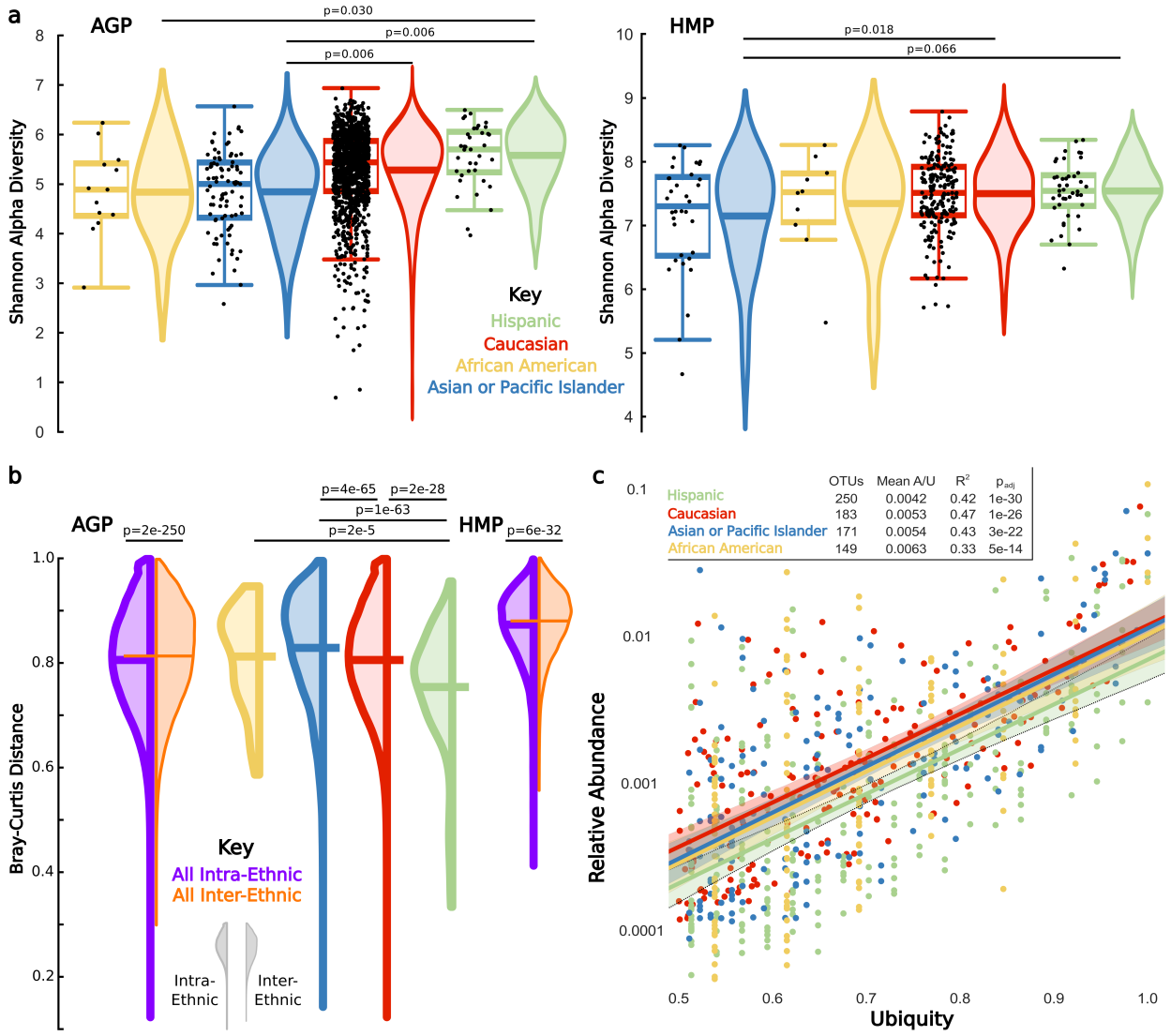


Fig 2. Ethnicity associates with diversity and composition of the gut microbiota. (A) Center lines of each boxplot depict the median by which ethnicities were ranked from low (left) to high (right); the lower and upper ends of each box represent the 25th and 75th quartiles respectively; whiskers denote the 1.5 interquartile range, and black dots represent individual samples. Lines in the middle of violin plots depict the mean, and p-values are Bonferroni corrected within each dataset. (B) Left extending violin plots represent intra-ethnic distances for each ethnicity, and right extending violin plots depict all inter-ethnic distances. Center lines depict the mean beta diversity. Significance bars above violin plots

150 depict Bonferroni corrected pairwise Mann-Whitney-U comparisons of the intra-intra- and
151 intra-inter-ethnic distances. (C) Within each ethnicity, OTUs shared by at least 50% of
152 samples. Colored lines represent a robust ordinary least squares regression within OTUs of
153 each ethnicity, shaded regions represent the 95% confidence interval, R^2 denotes the
154 regression correlation, the OTUs column indicates the number of OTUs with >50% ubiquity
155 for that ethnicity, Mean A/U is the average abundance/ubiquity ratio, and the p_{adj} is the
156 regression significance adjusted and Bonferroni corrected for the number of ethnicities.
157
158 If ethnicity impacts microbiota composition, pairwise beta diversity distances
159 (ranging from 0/completely dissimilar to 1/identical) will be greater between ethnicities
160 than within ethnicities. While average gut microbiota beta diversities across all individuals
161 are high (**Fig 2B**, Bray-Curtis=0.808), beta diversities between individuals of the same
162 ethnicity (intra-ethnic, Bray-Curtis=0.806) are subtly, but significantly, lower than those
163 between ethnicities in both the AGP (inter-ethnic, Bray-Curtis=0.814) and HMP datasets
164 (intra-ethnic, Bray-Curtis=0.870 versus inter-ethnic, Bray-Curtis=0.877). We confirm AGP
165 results by subsampling individuals from overrepresented ethnicities across beta metrics and
166 rarefaction depths (**S4A-4B Tables**). Finally, we repeat analyses across beta metrics and
167 rarefaction depths using only the average distance of each individual to all individuals from
168 the ethnicity to which they are compared (**S4C-4D Tables**).
169 Next, we explore inter-ethnic differences in the number of OTUs shared in at least
170 50% of individuals within an ethnicity, as the likelihood of detecting a biological signal is
171 improved in more abundant organisms relative to noise that may predominate in lower
172 abundance OTUs. Out of 5,591 OTUs in the total AGP dataset, 101 (1.8%) meet this ubiquity

173 cutoff in all ethnicities, and 293 (5.2%) OTUs meet the cutoff within at least one ethnicity.
174 Hispanics share the most ubiquitous OTUs and have the lowest average abundance/ubiquity
175 (A/U) ratio (**Fig 2C**), indicating stability whereby stability represents a more consistent
176 appearance of OTUs with lower abundance but higher ubiquity (32). This result potentially
177 explains their significantly lower intra-ethnic beta diversity distance and thus higher
178 microbial community overlap relative to the other ethnicities (**Fig 2B**). Comparisons in the
179 AGP between the higher sampled Hispanic, Caucasian, and Asian-Pacific Islander ethnicities
180 also reveal a trend wherein higher intra-ethnic community overlap (**Fig 2B**) parallels higher
181 numbers of ubiquitous OTUs (**Fig 2C**), higher Shannon Alpha diversity (**Fig 2A**), and higher
182 stability of ubiquitous OTUs as measured by the abundance/ubiquity (A/U) ratio (**Fig 2C**).
183 We next assess whether a single ethnicity disproportionately impacts total gut
184 microbiota distinguishability in the AGP by comparing ANOSIM results from the consensus
185 beta diversity distance matrix when each ethnicity is sequentially removed from the analysis
186 (**Fig 3A** and **S2E Table**). Distinguishability remains unchanged when the few African
187 Americans are removed, but is lost upon removal of Asian-Pacific Islanders or Caucasians,
188 likely reflecting their higher beta diversity distance from other ethnicities (**Fig 3A**). Notably,
189 removal of Hispanics increases distinguishability among the remaining ethnicities, which
190 may be due to higher degree of beta diversity overlap observed between Hispanics and other
191 ethnicities (**S4B Table**). Results conform across rarefaction depths and beta diversity
192 metrics (**S2F Table**), and pairwise combinations show strong distinguishability between
193 African Americans and Hispanics (ANOSIM, $R=0.234$, $p=0.005$), and Asian-Pacific Islanders
194 and Caucasians (ANOSIM, $R=0.157$, $p<0.001$).

Finally, to complement evaluation with ecological alpha and beta diversity we implement a random forest (RF) supervised learning algorithm to classify gut microbiota from genus level community profiles into their respective ethnicity. We build four one-versus-all binary classifiers to classify samples from each ethnicity compared to the rest, and use two different sampling approaches to train the models, Synthetic Minority Over-sampling Technique (SMOTE) (33) and down-sampling, for overcoming uneven representation of ethnicities in both the datasets (see Methods). Given that the area under the receiver operating characteristic (ROC) curve (or AUC) of a random guessing classifier is 0.5, the models classify each ethnicity fairly well (**Fig 3B**) with average AUCs across sampling techniques and datasets of 0.78 for Asian-Pacific Islanders, 0.76 for African Americans, 0.69 for Hispanics, and 0.70 for Caucasians. Ethnicity distinguishing RF taxa and out-of-bag error percentages appear in (**S2 Fig**).

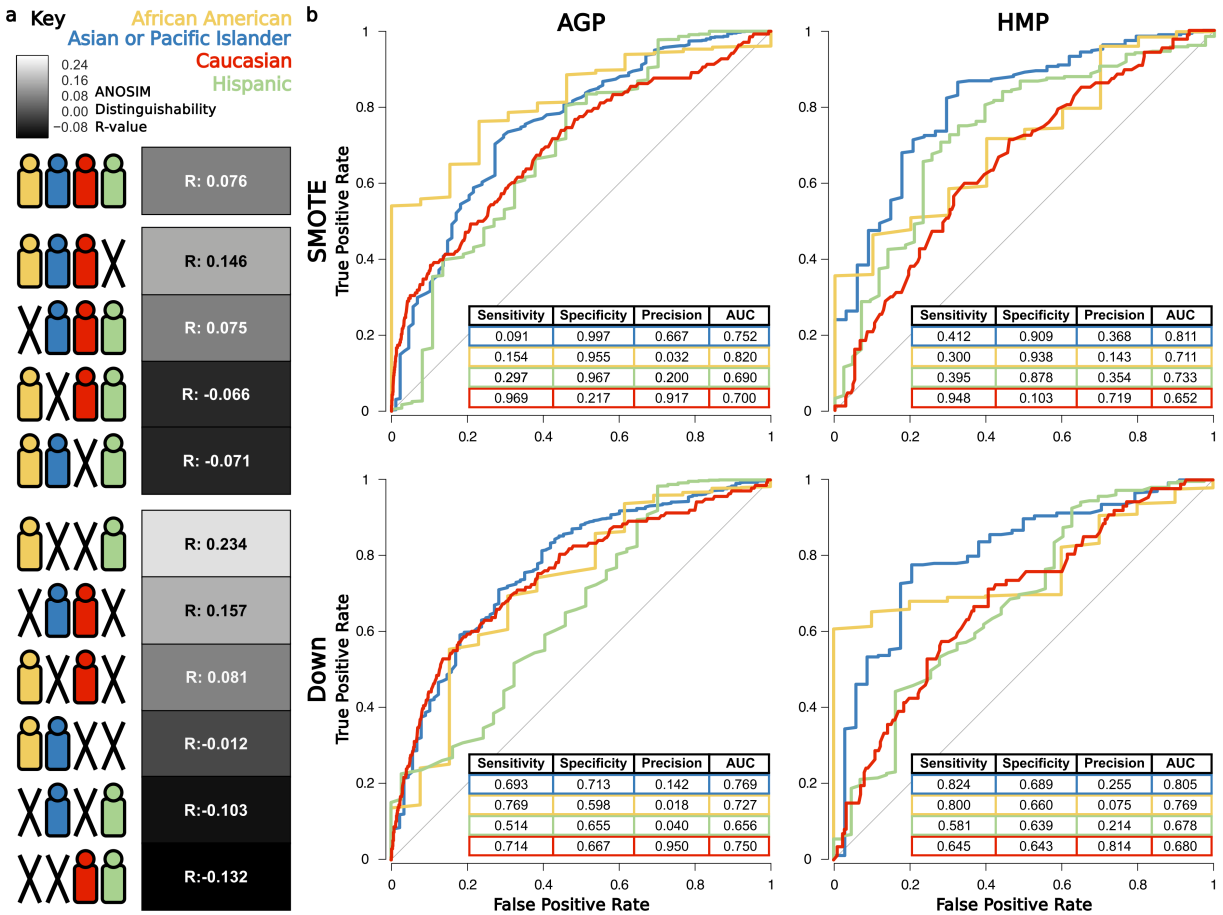


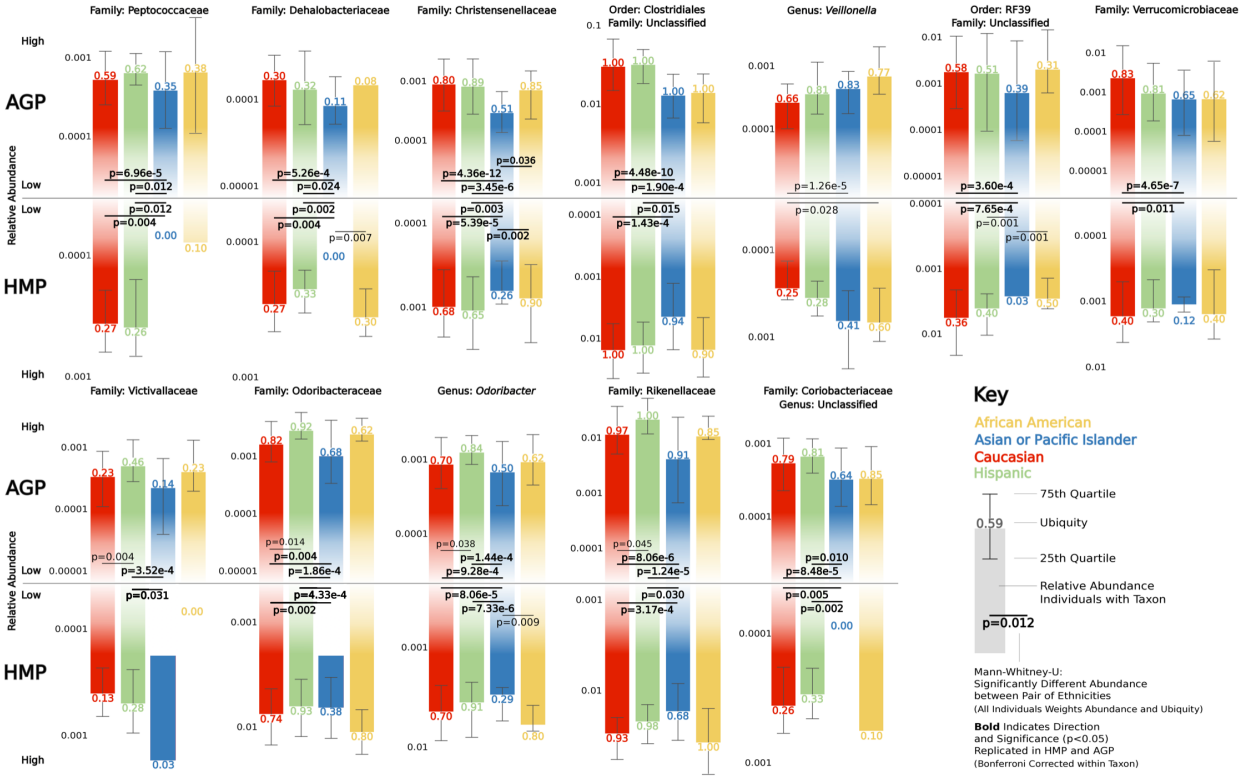
Fig 3. Microbiota distinguishability and classification ability across ethnicities. (A) ANOSIM distinguishability between all combinations of ethnicities. Symbols depict specific ethnicities included in the ANOSIM tests, and boxes denote the R-value as a heatmap, where white indicates increasing and black indicates decreasing distinguishability relative to the R-value with all ethnicities. (B) Average ROC curves (for 10-fold cross-validation) and prediction performance metrics for one-versus-all RF classifiers for each ethnicity, using SMOTE (33) and down subsampling approaches for training.

Recurrent taxon associations with ethnicity

217 Subtle to moderate ethnicity-associated differences in microbial communities may in
218 part be driven by differential abundance of certain microbial taxa. 16.2% (130/802) of the
219 AGP taxa and 20.6% (45/218) of HMP taxa across all classification levels (i.e. phylum to
220 genus, **S5 Table**) significantly vary in abundance across ethnicities (Kruskal-Wallis,
221 $p_{FDR}<0.05$). Between datasets, 19.2% (25/130) of the AGP and 55.6% (25/45) of the HMP
222 varying taxa replicate in the other dataset, representing a significantly greater degree of
223 overlap than would be expected by chance (ethnic permutation analysis of overlap, $p<0.001$
224 each taxonomic level and all taxonomic levels combined). The highest replication of taxa
225 varying by abundance occurs with 22.0% of families (9 significant in both datasets / 41
226 significantly varying families in either dataset), followed by genus with 13.4% (9 significant
227 in both datasets / 67 significantly varying genera in either dataset).

228 Among 18 reproducible taxa, we categorize 12 as taxonomically distinct (**Fig 4**) and
229 exclude 6 where nearly identical abundance profiles between family/genus taxonomy
230 overlap. Comparing relative abundance differences between pairs of ethnicities for these 12
231 taxa in the AGP reveals 30 significant differences, of which 20 replicate in the HMP ($p<0.05$,
232 Mann-Whitney-U). Intriguingly, all reproducible pairwise differences are a result of
233 decreases in Asian-Pacific Islanders (**Fig 4**). We also test taxon abundance and
234 presence/absence associations with ethnicity separately in the AGP using linear and logistic
235 regression models respectively, and we repeat the analysis while incorporating categorical
236 sex and continuous age and BMI as covariates (**S6 Table**). Clustering microbial families
237 based on their abundance correlation reveals two co-occurrence clusters: (i) a distinct
238 cluster of six Firmicutes and Tenericutes families in the HMP and (ii) an overlapping but
239 more diverse cluster of 20 families in the AGP (**S3 Fig**). Nine of the 12 taxa found to

recurrently vary in abundance across ethnicities are represented in these clusters (**Fig 4**), with four appearing in both clusters and the other five appearing either in or closely correlated with members of both clusters (**S3 Fig**). Furthermore, 90% (18/20) of families in the AGP cluster and 66% (4/6) of taxa in the HMP cluster significantly vary in abundance across ethnicities. We also found overlap for AGP and HMP datasets between taxa significantly varying in abundance across ethnicities (with FDR <0.05) and taxa in RF models with percentage importance greater than 50% for an ethnicity (**S2B Fig**). Taken together, these results establish general overlap of the most significant ethnicity-associated taxa between the these methods, reproducibility of microbial abundances that vary between ethnicities across datasets, and patterns of co-occurrence among these taxa which could suggest they are functionally linked.



252 **Fig 4. Ethnicity-associated taxa match between the HMP and AGP.** Barplots depict the
253 log10 transformed relative abundance for individuals possessing the respective taxon within
254 each ethnicity, ubiquity appears above (AGP) or below (HMP) bars, and the 25th and 75th
255 percentiles are shown with extending whiskers. Mann-Whitney-U tests evaluate differences
256 in abundance and ubiquity for all individuals between pairs of ethnicities; for example, the
257 direction of change in Victivallaceae is driven by ubiquity while abundance is higher for
258 those possessing the taxon. Significance values are Bonferroni corrected for the six tests
259 within each taxon and dataset, and bold p-values indicate that significance ($p<0.05$) and
260 direction of change replicate in the AGP and HMP.
261
262 ***Most heritable taxon of bacteria varies by ethnicity***
263 Identified as the most heritable taxon in the human gut (34, 35), the family
264 Christensenellaceae exhibits the second strongest significant difference in abundance across
265 ethnicities in both AGP and HMP datasets (**S5 Table**, Family: AGP, Kruskal-Wallis,
266 p_{FDR}=1.55e-9; HMP, Kruskal-Wallis, p_{FDR}=0.0019). Additionally, Christensenellaceae is
267 variable by sex and BMI (AGP: Sex, Kruskal-Wallis, p_{FDR}=1.22e-12; BMI, Kruskal-Wallis,
268 p_{FDR}=0.0020), and represents some of the strongest pairwise correlations with other taxa in
269 both co-occurrence clusters (**S3 Fig**). There is at least an eight-fold and two-fold reduction
270 in average Christensenellaceae abundance in Asian-Pacific Islanders relative to the other
271 ethnicities in the AGP and HMP respectively (**S5 Table**), and significance of all pairwise
272 comparisons in both datasets show reduced abundance in Asian-Pacific Islanders (**Fig 4**).
273 Christensenellaceae also occur among the top 10 most influential taxa for distinguishing
274 Asian-Pacific Islanders from other ethnicities using RF models for both AGP and HMP

datasets (**S2A Fig**). Abundance in individuals possessing Christensenellaceae and presence/absence across all individuals significantly associate with ethnicity (**S6 Table**, Abundance, Linear Regression, $p_{Bonferroni}$=0.006; Presence/Absence, Logistic Regression, $p_{Bonferroni}$=8.802e-6), but there was only a slight correlation between the taxon's relative abundance and BMI (**S4 Fig**). Confirming previous associations with lower BMI(36), we observe that AGP individuals with Christensenellaceae also have a lower BMI (Mean BMI, 23.7±4.3) than individuals without it (Mean BMI, 25.0±5.9; Mann-Whitney-U, p<0.001). This pattern is separately reflected in African Americans, Asian-Pacific Islanders, and Caucasians but not Hispanics (**Fig 5**), suggesting that each ethnicity may have different equilibria between the taxon's abundance and body weight.

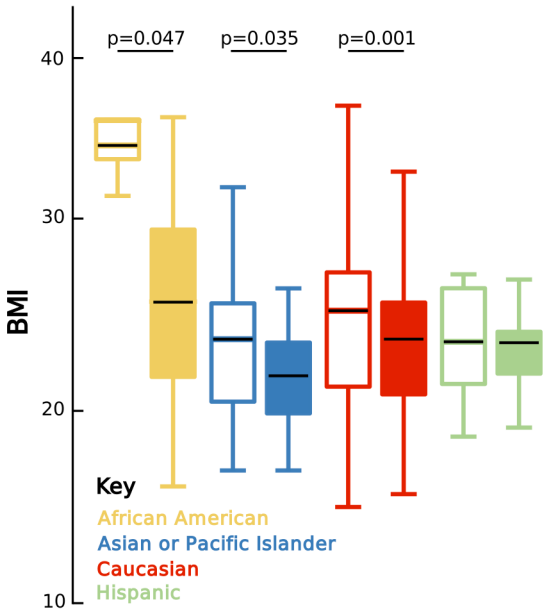


Fig 5. Christensenellaceae variably associate with BMI across ethnicities. Boxplots of BMI for individuals without (unfilled boxplots) and with (filled boxplots) Christensenellaceae. Significance was determined using one-tailed Mann-Whitney-U tests for lower continuous BMI values. Black lines indicate the mean relative abundance; the lower

290 and upper end of each box represent the 25th and 75th quartiles respectively; and whiskers
291 denote the 1.5 interquartile range.
292
293 ***Genetic- and ethnicity-associated taxa overlap***
294 Many factors associate with human ethnicity, including a small subset of population
295 specific genetic variants (estimated ~0.5% genome wide) that vary by biogeographical
296 ancestry (37, 38); self-declared ethnicity in the HMP is delineated by population genetic
297 structure (20). Here we investigate whether ethnicity-associated taxa overlap with (i) taxa
298 that have a significant population genetic heritability in humans (34, 35, 39, 40) and (ii) taxa
299 linked with human genetic variants in two large Genome-Wide Association Studies (GWAS)-
300 microbiota analyses (35, 40). All recurrent ethnicity-associated taxa except one were
301 heritable in at least one study, with seven replicating in three or more studies (**Table 1**).
302 Likewise, abundance differences in seven recurrent ethnicity-associated taxa demonstrate
303 significant GWAS associations with at least one variant in the human genome. Therefore, we
304 assess whether any genetic variants associated with differences in microbial abundance
305 exhibit significant rates of differentiation (F_{ST}) between 1,000 genomes superpopulations
306 (38). Out of 49 variants associated with ethnically varying taxa, 21 have higher F_{ST} values
307 between at least one pair of populations than that of 95% of other variants on the same
308 chromosome and across the genome; the F_{ST} values of five variants associated with
309 Clostridiaceae abundance rank above the top 99% (**S7 Table**). Since taxa that vary across
310 ethnicities exhibit lower abundance in Asian-Pacific Islanders, it is notable that the F_{ST} values
311 of 18 and 11 variant comparisons for East Asian and South Asian populations, respectively,
312 are above that of the 95% rate of differentiation threshold from African, American, or

313 European populations. Cautiously, the microbiota and 1,000 genomes datasets are not
314 drawn from the same individuals, and disentangling the role of genetic from social and
315 environmental factors will still require more controlled studies.
316

Recurrent Ethnicity-Associated Taxa	Heritability	Genetic Associations
Family: Peptococcaceae	0.1213 [A], 0.2154[C], 0.26[E]	rs143179968[E]
Family: Dehalobacteriaceae	0.6878[B], 0.3087[C]	
Family: Christensenellaceae	0.3819[A], 0.6170[B], 0.4230[C],	
Order: Clostridiales, Family: Unclassified	0.2914 [A], 0.4020[B], 0.1330[C]	*40 Genetic Variants[C]
Genus: *Veillonella*	0.1370 [A], 0.2168[D]	rs347941[C]
Order: RF39, Family: Unclassified	0.2341 [A], 0.6618[B], 0.3074[C]	rs4883972[C]
Family: Verrucomicrobiaceae	0.1257 [A], 0.5973[B], 0.1394[C]	
Family: Victivallaceae		
Family Odoribacteraceae	0.1389 [A], 0.1917[D], 0.34[E]	chr7:96414393[E], rs115795847[E]
Genus: *Odoribacter*	0.1916[D]	
Family: Rikenellaceae	0.1299[D], 0.29[E]	rs17098734[C], rs3909540[C], rs147600757[E]
Family: Coriobacteraceae, Genus: Unclassified	0.1364 [A], 0.2822[B], 0.1609[C]	rs9357092[E]

317

318 **Table 1. Most recurrent ethnicity-associated taxa are previously reported heritable**
319 **and genetically-associated taxa.** The table shows population genetic heritability estimates
320 and associated genetic variants for the 12 recurrent ethnically varying taxa. The minimum
321 heritability cutoff was chosen as >0.1, and only exactly overlapping taxonomies were
322 considered. Studies examined: [A]UKTwins (2014, 'A' measure of additive heritability in ACE
323 model) (34), [B]Yatsunenko (2014, 'A' measure of additive heritability in ACE model) (34),
324 [C]UKTwins (2016, 'A' measure of additive heritability in ACE model) (35), [D]Lim (2016, H2r
325 measure of polygenic heritability in SOLAR (41)) (39), [E]Turpin (2016, H2r measure of
326 polygenic heritability in SOLAR (41)). *indicates excessive variants were excluded from
327 table.
328

329 **Discussion:**

330 Many common diseases associate with microbiota composition and ethnicity, raising
331 the central hypothesis that microbiota differences between ethnicities can occasionally
332 serve as a mediator of health disparities. American's self-declared ethnicity can capture
333 socioeconomic, cultural, geographic, dietary and genetic diversity, and a similarly complex
334 array of interindividual and environmental factors influence total microbiota composition.
335 This complexity may result in challenges when attempting to recover consistent trends in
336 total gut microbiota differences between ethnicities. The challenges in turn emphasize the
337 importance of reproducibility, both through confirmation across analytical methods and
338 replication across study populations (15-17, 20, 27, 42). In order to robustly substantiate the
339 ethnicity-microbiota hypothesis, we evaluated recurrent associations between self-declared
340 ethnicity and variation in both total gut microbiota and specific taxa in healthy individuals.
341 Results provide hypotheses for examining specific members of the gut microbiota as
342 mediators of health disparities.

343 Our findings from two American datasets demonstrate that: (i) ethnicity consistently
344 captures gut microbiota with a slightly stronger effect size than other variables such as BMI,
345 age, and sex, (ii) ethnicity is moderately predictable from total gut microbiota differences,
346 and (iii) 12 taxa recurrently vary in abundance between the ethnicities, of which the majority
347 have been previously shown to associate with human genetic variation. Whether shaped
348 through socioeconomic, dietary, healthcare, genetic, or other ethnicity-related factors,
349 reproducibly varying taxa represent sources for novel hypotheses addressing health
350 disparities. For instance, the family Odoribacteraceae and genus *Odoribacter* are primary
351 butyrate producers in the gut, and they have been negatively associated to severe forms of
352 Crohn's disease and Ulcerative Colitis in association with reduced butyrate metabolism (43-

353 45). Asian-Pacific Islanders possess significantly less Odoribacteraceae and *Odoribacter* than
354 Hispanics and Caucasians in both datasets, and severity of Ulcerative Colitis upon hospital
355 admission has been shown to be significantly higher in Asian Americans (46). Considering
356 broader physiological roles, several ethnicity-associated taxa are primary gut anaerobic
357 fermenters and methanogens (47, 48), and associate with lower BMI and blood triglyceride
358 levels (36, 49). Indeed, Christensenellaceae, Odoribacteraceae, *Odoribacter*, and the class
359 Mollicutes containing RF39 negatively associate with metabolic syndrome and demonstrate
360 significant population genetic heritability in twins (39). Implications for health outcomes
361 warrant further investigation, but could be reflected by positive correlations of
362 Odoribacteraceae, *Odoribacter*, Coriobacteriaceae, Christensenellaceae, and the dominant
363 Verrucomicrobiaceae lineage *Akkermansia* with old age (50, 51). *Akkermansia* associations
364 with health and ethnicity in western populations may reflect recently arising dietary and
365 lifestyle effects on community composition, as this mucus consuming taxon is rarely
366 observed in more traditional cultures globally (23). Moreover, these findings raise the
367 importance of controlling for ethnicity in studies linking microbiota differences to disease
368 because associations between specific microbes and a disease could be confounded by
369 ethnicity of the study participants.

370 Based on correlations in individual taxon's abundance, a similar pattern of co-
371 occurrence previously identified as the 'Christensenellaceae Consortium' includes 11 of the
372 12 recurrent ethnically varying taxa (34), and members of this consortium associate with
373 genetic variation in the human formate oxidation gene *ALDH1L1*, which is a genetic risk
374 factor for stroke (35, 52, 53). Formate metabolism is a key step in the pathway reducing
375 carbon dioxide to methane (54, 55), and increased methane associates with increased

376 Rikenellaceae, Christensenellaceae, Odoribacteraceae and *Odoribacter* (56). Products of
377 methanogenic fermentation pathways include short chain fatty acids such as butyrate, which
378 through reduction of pro-inflammatory cytokines is linked to cancer cell apoptosis and
379 reduced risk of colorectal cancer (57, 58). Asian Americans are the only ethnic group where
380 cancer surpasses heart disease as the leading cause of death, and over 70% of Asian
381 Americans were born overseas, which can affect assimilation into western lifestyles, leading
382 to reduced access to healthcare and screening, and proper medical education (57, 59-61).
383 Preliminary results from other groups suggest that the gut microbiome of Southeast Asian
384 immigrants changes after migration to the United States (Dan Knights, personal
385 communication). Indeed, as countries in Asia shift toward a more western lifestyle, the
386 incidence of cancers, particularly gastrointestinal and colorectal cancers, are increasing
387 rapidly, possibly indicating incompatibilities between traditionally harbored microbiota and
388 western lifestyles (62-65). Asian Americans have higher rates of type 2 diabetes and
389 pathogenic infections than Caucasians (66), and two metagenomic functions enriched in
390 control versus type 2 diabetes cases appear to be largely conferred by cluster-associated
391 butyrate-producing and motility-inducing Verrucomicrobiaceae and Clostridia taxa reduced
392 in abundance among AGP and HMP Asian-Pacific Islanders (11). Both induction of cell
393 motility and butyrate promotion of mucin integrity can protect against pathogenic
394 colonization and associate with microbial community changes (11, 58, 67). Levels of cell
395 motility and butyrate are key factors suspected to underlie a range of health disparities
396 including inflammatory bowel disease, arthritis, and type 2 diabetes (11, 68-70). Patterns of
397 ethnically varying taxa across ethnicities could result from many factors including varying
398 diets, environmental exposures, sociocultural influences, human genetic variation and

399 others. However, regardless of the mechanisms dictating assembly, these results suggest
400 there is a reproducible, co-occurring group of taxa linked by similar metabolic processes
401 known to promote homeostasis.

402 The utility of this work is establishing a framework for studying ethnicity-associated
403 taxa and hypotheses of how changes in abundance or presence of these taxa may or may not
404 shape health disparities, many of which also have genetic components. Differing in allele
405 frequency across three population comparisons and associated with the abundance of
406 Clostridiales, the genetic variant rs7587067 has a significantly higher frequency in African
407 (Minor Allele Frequency (MAF)=0.802) versus East Asian (MAF=0.190, F_{ST}=0.54,
408 Chromosome=98.7%, Genome-Wide=98.9%), admixed American (MAF=0.278, F_{ST}=0.44,
409 Chromosome=99.0%, Genome-Wide=99.1%), and European populations (MAF=0.267,
410 F_{ST}=0.45, Chromosome=98.7.3%, Genome-Wide=98.7%). This intronic variant for the gene
411 *HECW2* is a known eQTL (GTeX, eQTL Effect Size=-0.18, p=7.4e-5) (71, 72), and *HECW2*
412 encodes a ubiquitin ligase linked to enteric gastrointestinal nervous system function through
413 maintenance of endothelial lining of blood vessels (73, 74). Knockout of *HECW2* in mice
414 reduced enteric neuron networks and gut motility, and patients with Hirschsprung's disease
415 have diminished localization of *HECW2* to regions affected by loss of neurons and colon
416 blockage when compared to other regions of their own colon and healthy individuals (75).
417 Hirschsprung's disease presenting as full colon blockage is rare and has not undergone
418 targeted examination as a health disparity, however a possible hypothesis is that lower
419 penetrance of the disease in individuals with the risk allele at rs7587067 could lead to
420 subtler effects on gut motility resulting in Clostridiales abundance differences.

421 Despite the intrigue of connecting the human genome, microbiota and disease
422 phenotypes, evaluating such hypotheses will require more holistic approaches including
423 incorporating metagenomics and metabolomics to identify whether enzymes or metabolic
424 functions reproducibly vary across ethnicities, as well as direct functional studies in model
425 systems to understand if correlation is truly driven by causation. Further limitations should
426 also be considered, including recruitment biases for the AGP versus HMP, variation in sample
427 processing and OTU clustering, and uneven sampling which could only be addressed with
428 down sampling of over-represented ethnicities. Still, despite these confounders care was
429 taken to demonstrate the reproducibility of results across statistical methods, ecological
430 metrics, rarefaction depths, and study populations. Summarily, this work suggests that
431 abundance differences of specific taxa, rather than whole communities, may represent the
432 most reliable ethnic signatures in the gut microbiota. A reproducible co-occurring subset of
433 these taxa link to a variety of overlapping metabolic processes and health disparities, and
434 contain the most reproducibly heritable taxon, Christensenellaceae. Moreover, a majority of
435 the microbial taxa associated with ethnicity are also heritable and genetically-associated
436 taxa, suggesting there is a possible connection between ethnicity and genetic patterns of
437 biogeographical ancestry that may play a role in shaping these taxa. Our results emphasize
438 the importance of sampling ethnically diverse populations of healthy individuals in order to
439 discover and replicate ethnicity signatures in the human gut microbiota, and they highlight
440 a need to account for ethnic variation as a potential confounding factor in studies linking
441 microbiota differences to disease. Further reinforcement of these results may lead to
442 generalizations about microbiota assembly and even consideration of specific taxa as
443 potential mediators or treatments of health disparities.

444 **Materials and Methods:**

445 *Data Acquisition*

446 AGP data was obtained from the project FTP repository located at
447 *ftp://ftp.microbio.me/AmericanGut/*. AGP data generation and processing prior to analysis
448 can be found at: *https://github.com/biocore/American-Gut/tree/master/ipynb/primary-*
449 *processing*. All analyses utilized the rounds-1-25 dataset which was released on March 4,
450 2016. Throughout all analyses, QIIME v1.9.0 was used in an Anaconda environment
451 [*https://continuum.io*] for all script calls, custom scripts and notebooks were run in the
452 QIIME 2 Anaconda environment with python version 3.5.2, and plots were post-processed
453 using Inkscape [*https://inkscape.org/en/*] (76). Ethnicity used in this study was self-declared
454 by AGP study participants as one of four groups: African American, Asian or Pacific Islander
455 (Asian-Pacific Islander), Caucasian, or Hispanic. Sex was self-declared as either male, female,
456 or other. Age was self-declared as a continuous integer of years old, and age categories
457 defined by the AGP by decade (i.e. 20's, 30's…) were used in this study. BMI was self-declared
458 as an integer, and BMI categories defined by AGP of underweight, healthy, overweight, and
459 obese were utilized. A total of 31 categorical metadata factors were assessed for structuring
460 across ethnicities with a two proportion Z test between pairs of ethnicities using a custom
461 python script (**S1 Table** additional sheets). The p-values were Bonferroni corrected within
462 each metadata factor for the number of pairwise ethnic comparisons. 97% Operational
463 Taxonomic Units (OTUs) generated for each dataset are utilized throughout to maintain
464 consistency with other published literature, however microbial taxonomy of the HMP is
465 reassigned using the Greengenes reference database (77). Communities characterized with
466 16S rDNA sequencing of variable region four followed an identical processing pipeline for all

467 samples, which was developed and optimized for the Earth Microbiome Project (78). HMP
468 16S rDNA data processed using QIIME for variable regions 3-5 was obtained from
469 *http://hmpdacc.org/HMQCP/*. Demographic info for individual HMP participants was
470 obtained through dbGaP restricted access to study phs000228.v2.p1, with dbGaP approval
471 granted to SRB and non-human subjects determination IRB161231 granted by Vanderbilt
472 University. Ethnicity and sex were assigned to subjects based on self-declared values, with
473 individuals selecting multiple ethnicities being removed unless they primarily responded as
474 Hispanic, while categorical age and BMI were established from continuous values using the
475 same criteria for assignment as in AGP. The HMP Amerindian population was removed due
476 to severe under-representation. This filtered HMP table was used for community level
477 analyses (ANOSIM, Alpha Diversity, beta intra-inter), however to allow comparison with the
478 AGP dataset, community subset analyses (co-occurrence, abundance correlation, etc…) were
479 performed with taxonomic assignments in QIIME using the UCLUST method with the
480 GreenGenes_13_5 reference.
481
482 *Quality Control*
483 AGP quality control was performed in Stata v12 (StataCorp, 2011) using available
484 metadata to remove samples (Raw N=9,475): with BMI more than 60 (-988 [8,487]) or less
485 than 10 (-68 [8,419]), missing age (-661 [7,758]), with age greater than 55 years old (-2,777
486 [4,981]) or less than 18 years old (-582 [4,399]), and blank samples or those not appearing
487 in the mapping file (-482 [3,917]), with unknown ethnicity or declared as other (-131
488 [3786]), not declared as a fecal origin (-2,002 [1784]), with unknown sex or declared as other
489 (-98 [1686]), or located outside of the United States (-209 [1477]). No HMP individuals were

490 missing key metadata or had other reasons for exclusion (-0[298]). Final community quality
491 control for both AGP and HMP was performed by filtering OTUs with less than 10 sequences
492 and removing samples with less than 1,000 sequences (AGP, -102 [1375]; HMP, -0 [298]). All
493 analyses used 97% OTUs generated by the AGP or HMP, and unless otherwise noted, results
494 represent Bray-Curtis beta diversity and Shannon alpha diversity at a rarefaction depth of
495 1,000 counts per sample.
496
497 *ANOSIM, PERMANOVA, and BioEnv Distinguishability*
498 The ANOSIM test was performed with 9,999 repetitions on each rarefied table within
499 a respective rarefaction depth and beta diversity metric (**Fig 1 & S2A-B Table**), with R-
500 values and p-values averaged across the rarefactions. Consensus beta diversity matrices
501 were calculated as the average distances across the 100 rarefied matrices for each beta
502 diversity metric and depth. Consensus distance matrices were randomly subsampled ten
503 times for subset number of individuals from each ethnic group with more than that subset
504 number prior to ANOSIM analysis with 9,999 repetitions, and the results were averaged
505 evaluating the effects of more even representations for each ethnicity (**S2C Table**).
506 Consensus distance matrices had each ethnicity and pair of ethnicities removed prior to
507 ANOSIM analysis with 9,999 repetitions, evaluating the distinguishability conferred by
508 inclusion of each ethnicity (**Fig 3A**, **S2F Table**). Significance was not corrected for the
509 number of tests to allow comparisons between results of different analyses, metrics, and
510 depths. PERMANOVA analyses were run using the R language implementation in the Vegan
511 package (79), with data handled in a custom R script using the Phyloseq package (80).
512 Categorical variables were used to evaluate the PERMANOVA equation (Beta-Diversity

513 Distance Matrix ~ Ethnicity + Age + Sex + BMI) using 999 permutations to evaluate
514 significance, and the R and p values were averaged across 10 rarefactions (**S2D Table**). The
515 BioEnv test, or BEST test, was adapted to allow evaluation of the correlation and significance
516 between beta diversity distance matrices and age, sex, BMI, and ethnicity simultaneously
517 (**S2E Table**) (29). At each rarefaction depth and beta diversity metric the consensus distance
518 matrix was evaluated for its correlation with the centered and scaled Euclidian distance
519 matrix of individuals continuous age and BMI, and categorical ethnicity and sex encoded
520 using patsy (same methodology as original test)[*https://patsy.readthedocs.io/en/latest/#*].
521 The test was adapted to calculate significance for a variable of interest by comparing how
522 often the degree of correlation with all metadata variables (age, sex, BMI, ethnicity) was
523 higher than the correlation when the variable of interest was randomly shuffled between
524 samples 1,000 times.
525
526 *Alpha Diversity*
527 Alpha diversity metrics (Shannon, Simpson, Equitability, Chao1, Observed OTUs)
528 were computed for each rarefied table (QIIME: alpha_diversity.py), and results were collated
529 and averaged for each sample across the tables (QIIME: collate_alpha.py). Pairwise
530 nonparametric t-tests using Monte Carlo permutations evaluated alpha diversity differences
531 between the ethnicities with Bonferroni correction for the number of comparisons (**Fig 2A**,
532 **S3 Table**, QIIME: compare_alpha_diversity.py). A Kruskal-Wallis test implemented in python
533 was used to detect significant differences across all ethnicities.
534
535 *Beta Diversity*

536 Each consensus beta diversity distance matrix had distances organized based on
537 whether they represented individuals of the same ethnic group, or were between individuals
538 of different ethnic groups. All values indicate that all pairwise distances between all
539 individuals were used (**Fig 2B**, **S4A-B Table**), mean values indicate that for each individual
540 their average distance to all individuals in the comparison group was used as a single point
541 to assess pseudo-inflation (**S4C-D Table**). A Kruskal-Wallis test was used to calculate
542 significant differences in intra-ethnic distances across all ethnicities. Pairwise Mann-
543 Whitney-U tests were calculated between each pair of intra-ethnic distance comparisons,
544 along with intra-versus-inter ethnic distance comparisons. Significance was Bonferroni
545 corrected within the number of intra-intra-ethnic and intra-inter-ethnic distance groups
546 compared, with violin plots of intra- and inter-ethnic beta diversity distances generated for
547 each comparison.
548
549 *Random Forest*

550 RF models were implemented using taxa summarized at genus level, which
551 performed better compared to RF models using OTUs as features, both in terms of
552 classification accuracy and computational time. We first rarefied OTU tables at sequence
553 depth of 10,000 (using R v3.3.3 package *vegan's* rrarefy() function) and then summarized
554 rarefied OTUs at genus-level (or lower characterized level if genus was uncharacterized for
555 an OTU). We filtered for rare taxa by removing taxa present in fewer than half of the number
556 of samples in rarest ethnicity (i.e. fewer than 10/2 = 5 samples in HMP and 13/2 = 6 (rounded
557 down) in AGP), retaining 85 distinct taxa in HMP dataset and 322 distinct taxa in AGP dataset
558 at genus level. The resulting taxa were normalized to relative abundance and arcsin-sqrt

559 transformed before being used as features for the RF models. We initially built multi-class
560 RF model, but since the RF model is highly sensitive to the uneven representation of classes,
561 all samples were identified as the majority class, i.e. Caucasian. In order to even out the class
562 imbalance, we considered some sampling approaches, but most existing techniques for
563 improving classification performance on imbalanced datasets are designed for binary class
564 imbalanced datasets, and are not effective on datasets with multiple underrepresented
565 classes. Hence, we adopted the binary classification approach and built four one-versus-all
566 binary RF classifiers to classify samples from each ethnicity compared to the rest. 10-fold
567 cross-validation (using R package *caret* (81)) was performed using ROC as the metric for
568 selecting optimal model. The performance metrics and ROC curves were averaged across the
569 10 folds (**Fig 3B**). Without any sampling during training the classifiers, most samples were
570 identified as the majority class, i.e. the Caucasian, by all four one-versus-all RF classifiers. In
571 order to overcome this imbalance in class representation, we applied two sampling
572 techniques inside cross-validation: i) down-sampling, and ii) Synthetic Minority Over-
573 sampling Technique (or SMOTE) (33). In the down-sampling approach, the majority class is
574 down-sampled by random removal of instances from the majority class. In the SMOTE
575 approach, the majority class is down-sampled and synthetic samples from the minority class
576 are generated based on k-nearest neighbors technique (33). Note, the sampling was
577 performed inside cross-validation on training set, while the test was performed on
578 unbalanced held-out test set in each fold. In comparison to a no-sampling approach, which
579 classified most samples as the majority class, i.e., Caucasians, our sampling-based approach
580 leads to improved sensitivity for classification of minority classes on unbalanced test sets.
581 Nevertheless, the most accurate prediction remains for the inclusion in the majority class.

582 The ROC curves and performance metrics table in **Fig 3B** show the sensitivity-specificity
583 tradeoff and classification performance for one-versus-all classifier for each ethnicity for
584 both the sampling techniques applied on both the datasets. For both the datasets, down-
585 sampling shows higher sensitivity and lower specificity and precision for minority classes
586 (i.e. African Americans, Asian-Pacific Islanders and Hispanics) compared to SMOTE.
587 However, for the majority class (i.e. Caucasian), down-sampling lowers the sensitivity and
588 increases the specificity and precision compared to SMOTE. The sensitivity-specificity
589 tradeoff, denoted by the area under the ROC curve (or AUC) is reduced for Hispanics in both
590 the datasets. The most important taxa with >50% importance for predicting an ethnicity
591 using RF model with SMOTE sampling approach are shown in **S2A Fig**. Among the 10 most
592 important taxa for each ethnicity, there are 9 taxa which overlap between the AGP and HMP
593 datasets (highlighted by the blue rectangular box); however, which ethnicity they best
594 distinguish varies between the two datasets. Within each dataset we highlighted taxa which
595 are distinguishing in RF models and have distinguishing differential abundance in **S2B Fig**,
596 reporting both the FDR corrected significance for Kruskal-Wallis tests of differential
597 abundance, and the percent importance for the most distinguished ethnicity of each in RF
598 models. We also report out-of-bag errors for the final RF classifier that was built using the
599 optimal model parameters obtained from cross-validation approach corresponding to each
600 ethnicity and sampling procedure for both AGP and HMP datasets in **S2C Fig**.

601

602 *Taxon Associations*

603 Taxon differential abundance across categorical metadata groups was performed in
604 QIIME (QIIME: group_significance.py, **S5 Table**) to examine whether observation counts (i.e.

605 OTUs and microbial taxon) are significantly different between groups within a metadata
606 category (i.e. ethnicity, sex, BMI, age). The OTU table prior to final community quality control
607 was collapsed at each taxonomic level (i.e. Phylum – Genus; QIIME: collapse_taxonomy.py),
608 with counts representing the relative abundance of each microbial taxon. Differences in the
609 mean abundance of taxa between ethnicities were calculated using Kruskal-Wallis
610 nonparametric statistical tests. P-values are provided alongside false discovery rate and
611 Bonferroni corrected P-values, and taxon were ranked from most to least significant. Results
612 were collated into excel tables by taxonomic level and metadata category being examined,
613 with significant (false discovery rate and Bonferroni P-value < 0.05) highlighted in orange,
614 and taxa that were false discovery rate significant in both datasets were colored red. The
615 Fisher's exact test for the overlap of number of significant taxa between datasets was run at
616 the online portal (http://vassarstats.net/tab2x2.html), with the expected overlap calculated
617 as 5% of the number of significant taxa at all levels within the respective dataset, and the
618 observed 25 taxa that overlapped in our analysis. The permutation analysis was performed
619 by comparing the number of significant taxa (**S5 Table**, $p_{FDR}<0.05$) overlapping between the
620 AGP and HMP to the number overlapping when the Kruskal-Wallis test was performed 1,000
621 times with ethnicity randomly permuted. In 1/1000 runs there was one significant taxon
622 overlapping at the family level, and one in 3/1000 permutations at the genus level, with no
623 significant taxa overlapping in any repetitions at higher taxonomic levels. The 12 families
624 and genera that were significantly different were evaluated to not be taxonomically distinct
625 if their abundances across ethnicities at each level represented at least 82-100% (nearly all
626 >95%) of the overlapping taxonomic level, and the genera was used if classified, and family
627 level used if genera was unclassified (g__). Average relative abundances on a log10 scale

628 among individuals possessing the taxon were extracted for each taxon within each ethnicity,
629 and the abundance for 12 families and genera were made into barchart figures (**Fig 4**). The
630 external whisker (AGP above, HMP below) depict the 75th quartile of abundance, and the
631 internal whisker depicts the 25th quartile. Pairwise Mann-Whitney-U tests were performed
632 between each pair of ethnicities using microbial abundances among all individuals, and were
633 Bonferroni corrected for the six comparisons within each taxon and dataset. Bonferroni
634 significant P-values are shown in the figure, and shown in bold if significance and direction
635 of change replicate in both datasets. Ubiquity shown above or below each bar was calculated
636 as the number of individuals in which that taxon was detected within the respective
637 ethnicity. Additional confirmation of ethnically varying abundance was also performed at
638 each taxonomic level (**S6 Table**), where the correlation of continuous age and BMI along
639 with categorically coded sex and ethnicity were simultaneously measured against the log 10
640 transformed relative abundance of each taxon among individuals possessing it using linear
641 regression (**S6 Table** - Abundance), and against the presence or absence of the taxon in all
642 individuals with logistic regression (**S6 Table** - Presence Absence). Significance is presented
643 for the models each with ethnicity alone, and with all metadata factors included (age, sex,
644 BMI), alongside Bonferroni corrected p-values, and individual effects of each metadata
645 factor.
646
647 Co-Occurrence Analysis
648 Bacterial taxonomy was collapsed at the family level, Spearman correlation was
649 calculated between each pair of families using SciPy (82), and clustermaps were generated
650 using seaborn (**S3 Fig**), and ethnic associations were drawn from **S5 Table**. Correlations

651 were masked where Bonferroni corrected Spearman p-values were >0.05, and clusters were
652 identified as the most prominent (strongest correlations) and abundance enriched.
653 Enrichment of ethnic association was evaluated by measuring the Mann-Whitney-U of
654 cluster families ethnic associations (p-values, **S5 Table**) compared to the ethnic associations
655 of non-cluster taxa. Cluster associated families were identified as having at least three
656 significant correlations with families within the cluster.
657
658 Christensenellaceae Analysis
659 The abundance of the family Christensenellaceae was input as relative abundance
660 across all individuals from the family level taxonomic table. Individuals were subset based
661 on the presence/absence of Christensenellaceae and BMIs were compared using a one tailed
662 Mann-Whitney-U test, then each was further subset by ethnicity and BMI compared using
663 one tailed Mann-Whitney-U tests and boxplots within each ethnicity (**Fig 5**).
664
665 Genetically Associated, Heritable, and Correlated Taxa Analysis
666 Genetically associated taxa from population heritability studies (34, 35, 39, 40) with
667 a minimum heritability (A in ACE models or H2r) >0.1, and from GWAS studies (35, 40) were
668 examined for exact taxonomic overlap with our 12 ethnically-associated taxa. The 42 genetic
669 variants associated with Unclassified Clostridiales are: rs16845116, rs586749, rs7527642,
670 rs10221827, rs5754822, rs4968435, rs17170765, rs1760889, rs6933411, rs2830259,
671 rs7318523, rs17763551, rs2248020, rs1278911, rs185902, rs2505338, rs6999713,
672 rs5997791, rs7236263, rs10484857, rs9938742, rs1125819, rs4699323, rs641527,
673 rs7302174, rs2007084, rs2293702, rs9350764, rs2170226, rs2273623, rs9321334,

674 rs6542797, rs9397927, rs2269706, rs4717021, rs7499858, rs10148020, rs7524581,
675 rs11733214, rs7587067 from (35). These 40 variants along with variants in **Table 1** except
676 for chr7:96414393 (total=49) were then assessed in 1,000 Genomes individuals for
677 significant differentiation across superpopulations (38). The 1,000 Genomes VCF files were
678 downloaded (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/), and variants
679 with a minor allele frequency less than 0.01 were removed with F_{ST} calculated between each
680 pair of superpopulations using vcftools (83). The East Asian versus South Asian F_{ST} rates
681 were not used in the analysis. A custom script was used to examine the F_{ST} for each of the 49
682 variants and compare to the F_{ST} of all variants on the same chromosome and all variants
683 genome-wide for that pair of populations, with percentile calculated and the number of
684 variants with a higher F_{ST} divided by the total number of variants. The eQTL value and
685 significance for rs7587067 were drawn from the GTEx database (72).
686
687 Data and Code Availability
688 Code, scripts, and data underlying figures are publicly available from the GitHub
689 repository [https://github.com/awbrooks19/microbiota_and_ethnicity]. Individual
690 metadata (age, sex, ethnicity…) for the Human Microbiome Project are held under restricted
691 access available through dbGaP application [NCBI - dbGaP, Human Microbiome Project,
692 https://www.ncbi.nlm.nih.gov/projects/gap/cgi-
693 bin/study.cgi?study_id=phs000228.v3.p1].
694
695 Acknowledgements

696 This work was supported by National Institutes of Health training grants
697 4T32GM08017810, 5T32GM08017809, and 5T32GM0817808 to AWB, the Vanderbilt Office
698 of Equity, Diversity and Inclusion to A.W.B. and S.R.B., the Vanderbilt Microbiome Initiative
699 to S.R.B., and the Alfred P. Sloan Foundation Fellowship to R.B.. The content is solely the
700 responsibility of the authors and does not necessarily represent the official views of the
701 National Institutes of Health. We thank the American Society of Microbiology for supporting
702 travel to present this work. We would also like to thank Tony Capra, David Samuels, Patrick
703 Abbot, Antonis Rokas, and other members of the Vanderbilt Genetics Institute and
704 Bordenstein Lab for input. The authors acknowledge the Minnesota Supercomputing
705 Institute (MSI) at the University of Minnesota and the Advanced Computing Center for
706 Research and Education (ACCRE) at Vanderbilt University for providing resources that
707 contributed to the research results reported within this paper.
708

709 Contributions

710 A.W.B., S.P., R.B., and S.R.B. conceived and designed the research. A.W.B. performed,
711 analyzed, and interpreted all experiments with the exception of the RF analysis planned and
712 performed by S.P. and R.B. S.R.B. supervised all experimental designs, data analysis, and data
713 interpretation. All authors participated in manuscript preparation, editing, and final
714 approval.

715

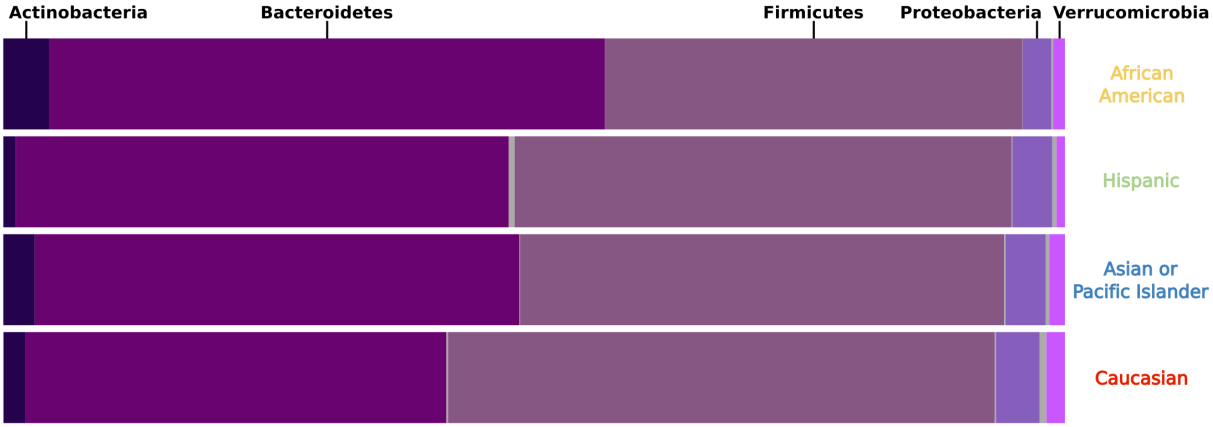
716 Competing Financial Interests

717 The authors declare no competing financial interests.

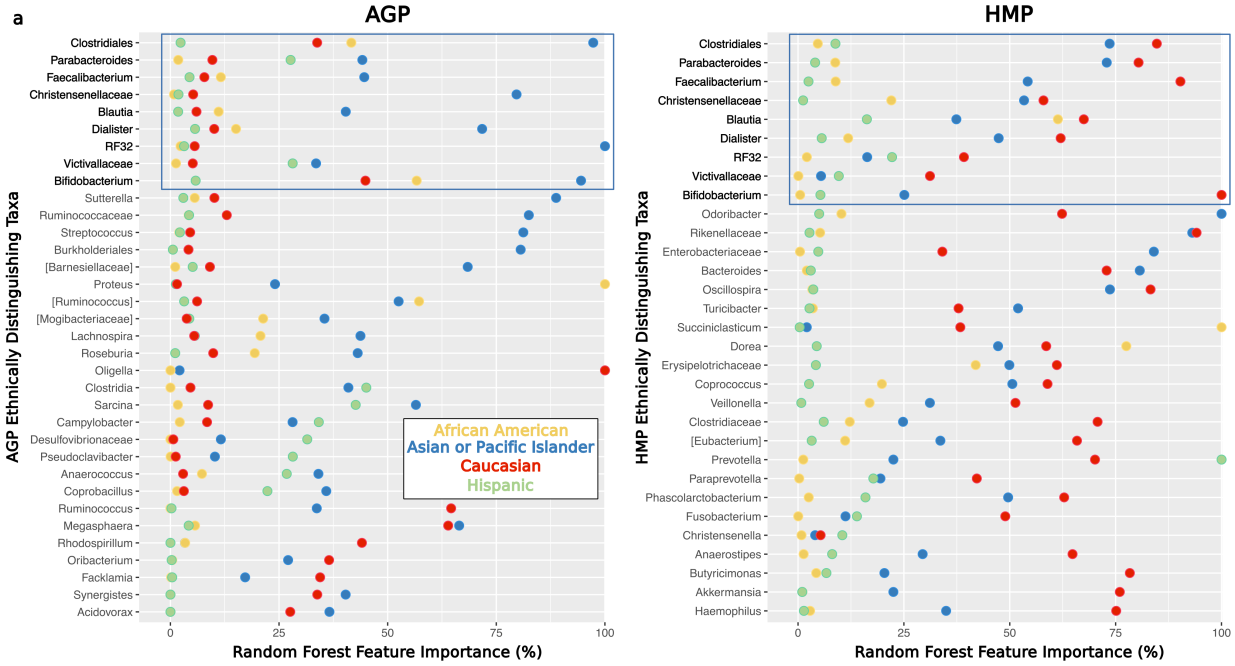
718

719

Supplementary Table/Figure Legends:



S1 Fig. The average relative abundance of dominant microbial phyla for each ethnicity.



b

AGP Taxonomic Lineage	Differential Abundance FDR	RF % Importance for Most Distinguished Ethnicity
Oligella	6.81E-19	100% for Caucasian
Bifidobacterium	1.63E-5	94.49% for Asian-Pacific Islander
Ruminococcaceae	3.92E-6	82.47% for Asian-Pacific Islander
Christensenellaceae	7.35E-9	79.64% for Asian-Pacific Islander
Dialister	0.0358	71.73% for Asian-Pacific Islander
Barnesiellaceae	0.0106	68.41% for Asian-Pacific Islander
Megasphaera	1.57E-5	66.45% for Asian-Pacific Islander

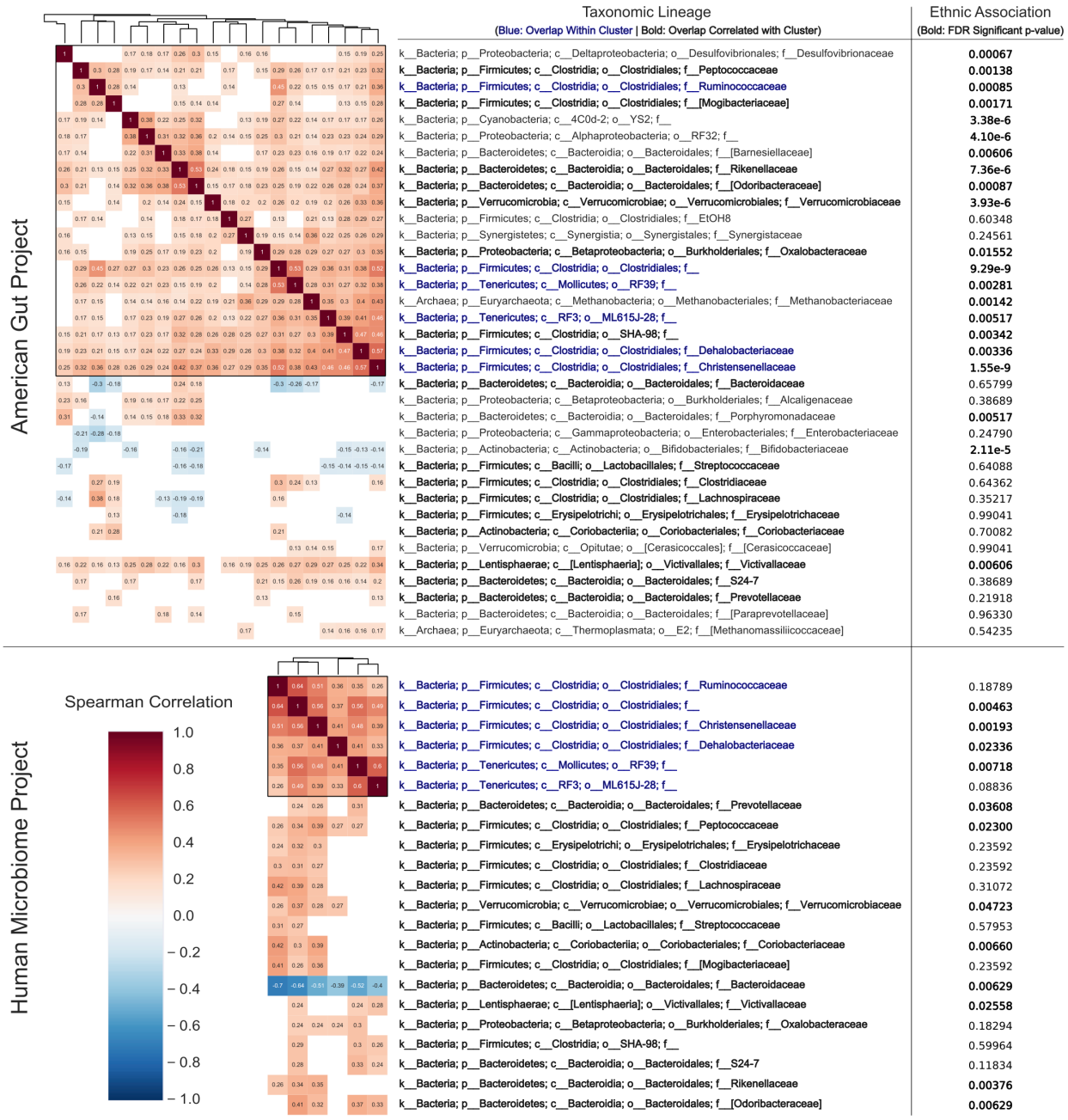
HMP Taxonomic Lineage	Differential Abundance FDR	RF % Importance for Most Distinguished Ethnicity
Odoribacter	0.0006	100% for Asian-Pacific Islander
Succiniclasticum	1.31E-7	100% for African American
Rikenellaceae	0.0042	94.09% for Caucasian
Bacteroides	0.0102	80.69% for Asian-Pacific Islander
Dorea	0.0362	77.50% for African American
Erysipelotrichaceae	0.0362	61.10% for Caucasian
Christensenellaceae	0.0027	57.95% for Caucasian
Veillonella	0.0432	51.32% for Caucasian

c

Sampling Method	African American	Asian or Pacific Islander	Caucasian	Hispanic
SMOTE	8.79%	11.69%	11.70%	8.49%
Down	42.31%	31.82%	29.35%	60.81%

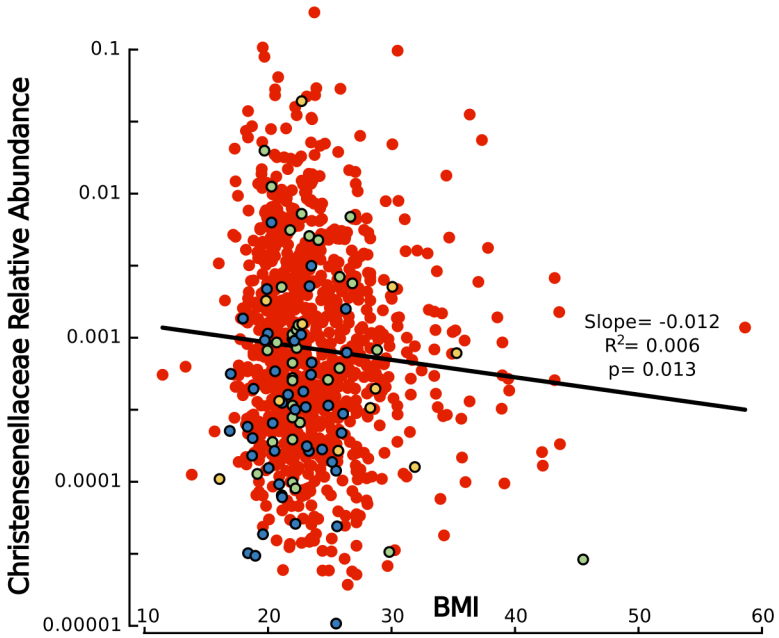
Sampling Method	African American	Asian or Pacific Islander	Caucasian	Hispanic
SMOTE	10.00%	7.56%	10.67%	12.29%
Down	40.00%	33.82%	38.51%	43.02%

724 **S2 Fig.** Summary of RF distinguishing taxa and out-of-bag error for each ethnicity. (A)
725 Importance of taxa for predicting each ethnicity using RF models with SMOTE sampling
726 approach are shown as percentage contributions, highlighted by color for each ethnicity.
727 Among the 10 most important taxa for each ethnicity, 9 overlap between the AGP and HMP
728 datasets (highlighted by the blue rectangular box), however which ethnicity they best
729 distinguish varies between the two datasets. (B) Taxa which are distinguishing in RF models
730 and have distinguishing differential abundance in **S5 Table**. The FDR corrected significance
731 for Kruskal-Wallis tests of differential abundance and the percent importance for the most
732 distinguished ethnicity of each in RF models are shown. (C) Out-of-bag error percentages for
733 the final RF classifier that was built using the optimal model parameters obtained from cross-
734 validation approach corresponding to each ethnicity and sampling procedure for both AGP
735 and HMP datasets.



S3 Fig. Abundance correlation of microbial families. Spearman correlation clustermaps of bacterial abundance for families in the AGP and HMP. Numbers within boxes depict the spearman correlation value with heatmap coloration from blue negative correlation (-1), white no correlation (0), to red positive correlation (1). Positions have been masked based on Bonferroni significance <0.05 for the total clustermap of all microbial families. Taxa

742 within boxes were identified as a highly correlated cluster, and taxa outside the boxes share
743 multiple correlations with those within the cluster. Blue taxonomic names indicate overlap
744 of taxa within boxes of both the AGP and HMP, while black indicate multiple correlations
745 with the clusters in both datasets. The ethnic association column depicts FDR corrected p-
746 values from Kruskal-Wallis tests in **S5 Table**, which are bolded if <0.05.



747

748 **S4 Fig. Correlation of BMI with Christensenellaceae abundance.** The relationship for
749 each individual between log10 transformed Christensenellaceae abundance on the y axis and
750 BMI on the x axis, with statistics slope, R^2, and p fit with a linear regression. Coloration of
751 each point indicates ethnicity: Yellow – African American; Blue – Asian-Pacific Islander;
752 Green – Hispanic; Red – Caucasian.

753

754 **S1 Table. Demographic information for the AGP.** Breakdown of age and BMI by sex and
755 ethnicity. Heatmaps were constructed within each statistic and category (bounded by black
756 box). The means for all sex and ethnic groups were used as the center (white), with higher

values indicated in red and lower in blue. HMP data is not shown because of data access restrictions on participant metadata, available through dbGaP application. Additional sheets depict proportions tests of ethnic structuring for 31 metadata factors, each on their own sheet.

S2 Table. Microbiota distinguishability by ethnicity, age, sex and BMI. (A) AGP and HMP ANOSIM distinguishability by ethnicity, age, sex, and BMI at a rarefaction depth of 1,000 and across four ecological metrics (more details in table). (B) AGP ANOSIM distinguishability by ethnicity, age, sex, and BMI at rarefaction depths of 1,000 and 10,000. (C) ANOSIM results for consensus distance matrix while subsampling the maximum number of individuals from each ethnic group. (D) BioEnv results of correlation between ethnicity, age, sex, and BMI together with outcome as multivariate beta diversity distance matrices [Distance Matrix = Ethnicity*x1 + Categorical Age*x2 + Categorical BMI*x3 + Sex*x4 + B]. (E) ANOSIM results for consensus distance matrix when each ethnicity and group of ethnicities are sequentially removed from the analysis.

S3 Table. Alpha diversity by ethnicity, age, sex and BMI. Alpha Diversity for Ethnicity, Age, Sex, and BMI across varying rarefaction depths and beta diversity metrics in AG (4A, 4C-E), and for ethnicity in the HMP (4B). Results are based on non-parametric permutation based t-tests, and p-values are Bonferroni corrected within each factor of interest, depth, and metric.

S4 Table. Comparison of beta diversity distances for within and between ethnicities. All values depicted are Mann-Whitney-U p-values. (A) All distances between pairs of individuals within each ethnicity were compared between ethnicities across rarefaction depths 1,000 and 10,000, four beta diversity metrics, and with while subsampling over-represented ethnicities. (B) All distances between pairs of individuals within and between each ethnicity were compared between ethnicities. (C) Mean distances between pairs of individuals within each ethnicity were compared between ethnicities. (D) Mean distances between pairs of individuals within and between each ethnicity were compared between ethnicities.

S5 Table. Taxa which are differentially abundant by ethnicity, sex, BMI, and age in the AGP and HMP. Kruskal-Wallis results for differential taxa abundance across metadata groupings, including FDR and Bonferroni corrected p-values, and taxa abundance averages within each group. Metadata factors and taxonomic levels are separated by excel tabs.

S6 Table. Taxa which are correlated with ethnicity, sex, BMI, and age in the AGP. Results of linear (Abundance) and logistic (Presence Absence) regression results for differential taxa abundance across metadata factors separated by taxonomic level. Columns in order indicate the taxon name, the number of individuals with non-zero abundance; then the p-value for ethnicity alone, the p-value Bonferroni corrected, the f-test statistic, and R^2; then the same values for the regression with ethnicity, age, sex, and BMI together; then the abundances in each ethnic group, and finally the p-values for each factor broken down.

802 **S7 Table. Genetic variants with taxa associations and detailed 1,000 Genomes**
803 **population differentiation rates (F_{ST}).** Variants in red indicate the variant has at least one
804 F_{ST} above the 95th percentile for high differentiation between at least one pair of populations.
805 Columns I-BU represent the values for calculating variant F_{ST} and percentiles. The first two
806 spaces indicate the two superpopulations being compared. F_{ST} indicates the rate of
807 differentiation for that variant between that pair of populations. Higher indicates the
808 number of variants genome-wide with a higher F_{ST}, and total indicates the total genome-wide
809 variants examined. The columns with chromosome indicate the number of variants with
810 higher F_{ST} and total variants on the same chromosome as the variant of interest. Percent
811 indicates the number of variants with a higher F_{ST} divided by the total number of variants.
812
813
814 **References:**

815 1. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, et al. A core gut
816 microbiome in obese and lean twins. Nature. 2009;457(7228):480-4.
817 2. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial
818 gene catalogue established by metagenomic sequencing. Nature. 2010;464(7285):59-65.
819 3. Huse SM, Ye Y, Zhou Y, Fodor AA. A core human microbiome as viewed through 16S
820 rRNA sequence clusters. PLoS One. 2012;7(6):e34242.
821 4. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen YY, Keilbaugh SA, et al. Linking long-term
822 dietary patterns with gut microbial enterotypes. Science. 2011;334(6052):105-8.
823 5. Muegge BD, Kuczynski J, Knights D, Clemente JC, González A, Fontana L, et al. Diet drives
824 convergence in gut microbiome functions across mammalian phylogeny and within humans.
825 Science. 2011;332:970-4.
826 6. Human Microbiome Project C. Structure, function and diversity of the healthy human
827 microbiome. Nature. 2012;486(7402):207-14.
828 7. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, et al. Diet
829 rapidly and reproducibly alters the human gut microbiome. Nature. 2013;505(7484):559-63.
830 8. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, et al.
831 Human gut microbiome viewed across age and geography. Nature. 2012;486(7402):222-7.
832 9. Davenport ER, Cusanovich DA, Michelini K, Barreiro LB, Ober C, Gilad Y. Genome-Wide
833 Association Studies of the Human Gut Microbiota. PLoS One. 2015;10(11):e0140301.

834 10. Fierera N, Hamadyc M, Lauberb CL, Knight R. The influence of sex, handedness, and
835 washing on the diversity of hand surface bacteria. Procedings of the National Academy of
836 Sciences.105(46).
837 11. Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, et al. A metagenome-wide association study of gut
838 microbiota in type 2 diabetes. Nature. 2012;490(7418):55-60.
839 12. Frank DN, Allison AL, Feldman RA, Boedeker EC, Harpaz N, Pace NR. Molecular-
840 phylogenetic characterization of microbial community imbalances in human inflammatory
841 bowel diseases. Procedings of the National Academy of Sciences. 2007(104):13780–5.
842 13. Walters WA, Xu Z, Knight R. Meta-analyses of human gut microbes associated with
843 obesity and IBD. FEBS Lett. 2014;588(22):4223-33.
844 14. Zackular JP, Baxter NT, Iverson KD, Sadler WD, Petrosino JF, Chen GY, et al. The gut
845 microbiome modulates colon tumorigenesis. MBio. 2013;4(6):e00692-13.
846 15. Mason MR, Nagaraja HN, Camerlengo T, Joshi V, Kumar PS. Deep sequencing identifies
847 ethnicity-specific bacterial signatures in the oral microbiome. PLoS One. 2013;8(10):e77287.
848 16. Ravela J, Gajera P, Abdob ZG, Schneiderc M, Koeniga SSK, McCullea SL, et al. Vaginal
849 microbiome of reproductive-age women. Procedings of the National Academy of Sciences.
850 2011;108:4680-7.
851 17. Fettweis JM, Brooks JP, Serrano MG, Sheth NU, Girerd PH, Edwards DJ, et al. Differences
852 in vaginal microbiome in African American women versus women of European ancestry.
853 Microbiology. 2014;160(Pt 10):2272-82.
854 18. Williams DR, Priest N, Anderson NB. Understanding associations among race,
855 socioeconomic status, and health: Patterns and prospects. Health Psychol. 2016;35(4):407-11.
856 19. Mersha TB, Abebe T. Self-reported race/ethnicity in the age of genomic research: its
857 potential impact on understanding health disparities. Human Genomics. 2015;9(1):1.
858 20. Kolde R, Franzosa EA, Rahnavard G, Hall AB, Vlamakis H, Stevens C, et al. Host genetic
859 variation and its microbiome interactions within the Human Microbiome Project. Genome Med.
860 2018;10(1):6.
861 21. Clemente JC PE, Blaser MJ, Sandhu K, Gao K, Wang B, Magda M, Hidalgo G, et al. The
862 microbiome of uncontacted Amerindians. Science Advances. 2015;3.
863 22. Rampelli S, Schnorr SL, Consolandi C, Turroni S, Severgnini M, Peano C, et al.
864 Metagenome Sequencing of the Hadza Hunter-Gatherer Gut Microbiota. Curr Biol.
865 2015;25(13):1682-93.
866 23. Smits SA, Leach J, Sonnenburg ED, Gonzalez CG, Lichtman JS, Reid G, et al. Seasonal
867 cycling in the gut microbiome of the Hadza hunter-gatherers of Tanzania. Science.
868 2017;357:802-6.
869 24. McDonald D, Birmingham A, Knight R. Context and the human microbiome.
870 Microbiome. 2015;3:52.
871 25. Blekhman R, Goodrich JK, Huang K, Sun Q, Bukowski R, Bell JT, et al. Host genetic
872 variation impacts microbiome composition across human body sites. Genome Biol.
873 2015;16:191.
874 26. Rothschild D, Weissbrod O, Barkan E, Kurilshikov A, Korem T, Zeevi D, et al. Environment
875 dominates over host genetics in shaping human gut microbiota. Nature. 2018;555(7695):210-5.

876 27. Deschasaux M, Bouter KE, Prodan A, Levin E, Groen AK, Herrema H, et al. Depicting the
877 composition of gut microbiota in a population with varied ethnic origins but shared geography.
878 Nat Med. 2018.
879 28. Clarke KR. Non-parametric multivariate analyses of changes in community structure.
880 Australian Journal of Ecology. 1993;18:117-43.
881 29. Clarke KR, Ainsworth M. A method of linking multivariate community structure to
882 environmental variables. Marine Ecology. 1993;92:205-19
883 .
884 30. Knights D, Costello EK, Knight R. Supervised classification of human microbiota. FEMS
885 Microbiol Rev. 2011;35(2):343-59.
886 31. Shannon CE. A mathematical theory of communication. Bell Syst Tech J. 1948;27:379-
887 423.
888 32. Hester ER, Barott KL, Nulton J, Vermeij MJ, Rohwer FL. Stable and sporadic symbiotic
889 communities of coral and algal holobionts. ISME J. 2016;10(5):1157-69.
890 33. Chawla NV, Bowyer KW, Hall LO and Kegelmeyer WP. SMOTE: Synthetic Minority Over-
891 sampling Technique. Journal of Artificial Intelligence Research. 2002;16:321-57.
892 34. Goodrich JK, Waters JL, Poole AC, Sutter JL, Koren O, Blekhman R, et al. Human genetics
893 shape the gut microbiome. Cell. 2014;159(4):789-99.
894 35. Goodrich JK, Davenport ER, Beaumont M, Jackson MA, Knight R, Ober C, et al. Genetic
895 Determinants of the Gut Microbiome in UK Twins. Cell Host Microbe. 2016;19(5):731-43.
896 36. Fu J, Bonder MJ, Cenit MC, Tigchelaar EF, Maatman A, Dekens JA, et al. The Gut
897 Microbiome Contributes to a Substantial Proportion of the Variation in Blood Lipids. Circ Res.
898 2015;117(9):817-24.
899 37. Pennisi E. Human Genetic Variation. Science. 2007;318(5858):1842-3.
900 38. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A
901 global reference for human genetic variation. Nature. 2015;526(7571):68-74.
902 39. Lim MY, You HJ, Yoon HS, Kwon B, Lee JY, Lee S, et al. The effect of heritability and host
903 genetics on the gut microbiota and metabolic syndrome. Gut. 2016.
904 40. Turpin W, Espin-Garcia O, Xu W, Silverberg MS, Kevans D, Smith MI, et al. Association of
905 host genome with intestinal microbial composition in a large healthy cohort. Nature Genetics.
906 2016;48(11):1413-7.
907 41. Almasy L, Blangero J. Multipoint quantitative-trait linkage analysis in general pedigrees.
908 Am J Hum Genet. 1998;62(5):1198-211.
909 42. Rothschild D, Weissbrod O, Barkan E, Korem T, Zeevi D, Costea PI, et al. Environmental
910 factors dominate over host genetics in shaping human gut microbiota composition. BioRxiv.
911 2017.
912 43. Morgan XC, Tickle TL, Sokol H, Gevers D, Huttenhower C. Dysfunction of the intestinal
913 microbiome in inflammatory bowel disease and treatment. Genome Biology. 2012;7(979).
914 44. Lewis JD, Chen EZ, Baldassano RN, Otley AR, Griffiths AM, Lee D, et al. Inflammation,
915 Antibiotics, and Diet as Environmental Stressors of the Gut Microbiome in Pediatric Crohn's
916 Disease. Cell Host Microbe. 2015;18(4):489-500.
917 45. Goker M, Gronow S, Zeytun A, Nolan M, Lucas S, Lapidus A, et al. Complete genome
918 sequence of Odoribacter splanchnicus type strain (1651/6). Stand Genomic Sci. 2011;4(2):200-
919 9.

920 46. Castaneda G, Liu B, Torres S, Bhuket T, Wong RJ. Race/Ethnicity-Specific Disparities in
921 the Severity of Disease at Presentation in Adults with Ulcerative Colitis: A Cross-Sectional Study.
922 Dig Dis Sci. 2017.
923 47. Boucias DG, Cai Y, Sun Y, Lietze VU, Sen R, Raychoudhury R, et al. The hindgut lumen
924 prokaryotic microbiota of the termite Reticulitermes flavipes and its responses to dietary
925 lignocellulose composition. Mol Ecol. 2013;22(7):1836-53.
926 48. LATHAM MJ, WOLIN MJ. Fermentation of Cellulose by Ruminococcus flavefaciens in the
927 Presence and Absence ofMethanobacterium ruminantium. Appl Environ Microbiol.
928 1977;34(3):297-301.
929 49. Falony G, Raes J. Population-level analysis of gut microbiome variation. Science.
930 2016;352(6285):560-4.
931 50. Biagi E, Franceschi C, Rampelli S, Severgnini M, Ostan R, Turroni S, et al. Gut Microbiota
932 and Extreme Longevity. Curr Biol. 2016;26(11):1480-5.
933 51. Thevaranjan N, Puchta A, Schulz C, Naidoo A, Szamosi JC, Verschoor CP, et al. Age-
934 Associated Microbial Dysbiosis Promotes Intestinal Permeability, Systemic Inflammation, and
935 Macrophage Dysfunction. Cell Host Microbe. 2017;21(4):455-66 e4.
936 52. Xie W, Wood AR, Lyssenko V, al. e. Genetic Variants Associated With Glycine
937 Metabolism and Their Role in Insulin Sensitivity and Type 2 Diabetes. Diabetes. 2013;62.
938 53. Williams SR, Yang Q, Chen F, Liu X, Keene KL, Jacques P, et al. Genome-wide meta-
939 analysis of homocysteine and methionine metabolism identifies five one carbon metabolism
940 loci and a novel association of ALDH1L1 with ischemic stroke. PLoS Genet.
941 2014;10(3):e1004214.
942 54. Petersen LM, Bautista EJ, Nguyen H, Hanson BM, Chen L, Lek SH, et al. Community
943 characteristics of the gut microbiomes of competitive cyclists. Microbiome. 2017;5(1):98.
944 55. Nakamura N, Lin HC, McSweeney CS, Mackie RI, Gaskins HR. Mechanisms of microbial
945 hydrogen disposal in the human colon and implications for health and disease. Annu Rev Food
946 Sci Technol. 2010;1:363-95.
947 56. Parthasarathy G, Chen J, Chen X, Chia N, O'Connor HM, Wolf PG, et al. Relationship
948 Between Microbiota of the Colonic Mucosa vs Feces and Symptoms, Colonic Transit, and
949 Methane Production in Female Patients With Chronic Constipation. Gastroenterology.
950 2016;150(2):367-79 e1.
951 57. Jackson CS, Oman M, Patel AM, Vega KJ. Health disparities in colorectal cancer among
952 racial and ethnic minorities in the United States. J Gastrointest Oncol. 2016;7(Suppl 1):S32-43.
953 58. Lopetuso LR, Scaldaferri F, Petito V, Gasbarrini A. Commensal Clostridia: leading players
954 in the maintenance of gut homeostasis. Gut Pathogens. 2013.
955 59. Sy DF. The Center for Asian Health Engages Communities in Research to Reduce Asian
956 American Health Disparities. US Department of Health & Human Services, National Institute on
957 Minority Health and Health Disparities.
958 60. Hwang H. Colorectal Cancer Screening among Asian Americans. Asian Pacific Journal of
959 Cancer Prevention. 2013;14(7):4025-32.
960 61. Oh KM, Kreps GL, Jun J. Colorectal Cancer Screening Knowledge, Beliefs, and Practices of
961 Korean Americans. American Journal of Health Behavior. 2013;37(3):381-94.
962 62. Sankaranarayanan R, Ramadas K, Qiao Y-l. Managing the changing burden of cancer in
963 Asia. BMC Medicine. 2014;12(3).

964 63. Pourhoseingholi MA. Increased burden of colorectal cancer in Asia. World J Gastrointest
965 Oncol. 2012;4(4):68-70.
966 64. Pourhoseingholi MA, Vahedi M, Baghestani AR. Burden of gastrointestinal cancer in
967 Asia; an overview. Gastroenterology and Hepatology. 2015.
968 65. Pourhoseingholi MA. Epidemiology and burden of colorectal cancer in Asia-Pacific
969 region: what shall we do now? Translational Gastrointestinal Cancer. 2014;3(4):169-73.
970 66. Report CHDaI. 2013.
971 67. Cao H, Liu X, An Y, Zhou G, Liu Y, Xu M, et al. Dysbiosis contributes to chronic
972 constipation development via regulation of serotonin transporter in the intestine. Sci Rep.
973 2017;7(1):10322.
974 68. Mosca A, Leclerc M, Hugot JP. Gut Microbiota Diversity and Human Diseases: Should We
975 Reintroduce Key Predators in Our Ecosystem? Front Microbiol. 2016;7:455.
976 69. Zhang X, Zhang D, Jia H, Feng Q, Wang D, Liang D, et al. The oral and gut microbiomes
977 are perturbed in rheumatoid arthritis and partly normalized after treatment. Nat Med.
978 2015;21(8):895-905.
979 70. Singh VP, Proctor SD, Willing BP. Koch's postulates, microbial dysbiosis and
980 inflammatory bowel disease. Clin Microbiol Infect. 2016;22(7):594-9.
981 71. Sherry ST WM, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI
982 database of genetic variation. Nucleic Acids Research. 2001;29(308).
983 72. Consortium GT. The Genotype-Tissue Expression (GTEx) project. Nat Genet.
984 2013;45(6):580-5.
985 73. Qiu X, Wei R, Li Y, Zhu Q, Xiong C, Chen Y, et al. NEDL2 regulates enteric nervous system
986 and kidney development in its Nedd8 ligase activity-dependent manner. Oncotarget.
987 2016;7(21).
988 74. Wei R, Qiu X, Wang S, Li Y, Wang Y, Lu K, et al. NEDL2 is an essential regulator of enteric
989 neural development and GDNF/Ret signaling. Cell Signal. 2015;27(3):578-86.
990 75. O'Donnell AM, Coyle D, Puri P. Decreased expression of NEDL2 in Hirschsprung's
991 disease. J Pediatr Surg. 2016;51(11):1839-42.
992 76. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al.
993 QIIME allows analysis of high-throughput community sequencing data. Nat Method.
994 2010;7:335-6.
995 77. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a
996 chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ
997 Microbiol. 2006;72(7):5069-72.
998 78. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, et al. Ultra-high-
999 throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J.
1000 2012;6(8):1621-4.
1001 79. Anderson MJ. A new method for non-parametric multivariate analysis of variance.
1002 Australian Journal of Ecology. 2001;26:32-46.
1003 80. McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis
1004 and graphics of microbiome census data. PLoS One. 2013;8:e61217.
1005 81. Kuhn M. A short introduction to the caret package. 2017.
1006 82. Jones E, Oliphant T, Peterson P. Open Source Scientific Tools for Python. 2001.

1007 83. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call
1008 format and VCFtools. Bioinformatics. 2011;27(15):2156-8.
1009