Short- and long-read metagenomics of South African gut microbiomes reveal a transitional composition and novel taxa

Fiona B. Tamburini[1], Dylan Maghini[1], Ovokeraye H. Oduaran[2], Ryan Brewster[3], Michaella R. Hulley[2,4], Venesa Sahibdeen[4], Shane A. Norris[5,6], Stephen Tollman[7,8], Kathleen Kahn[7,8], Ryan G. Wagner[7,8], Alisha N. Wade[7], Floidy Wafawanaka[7], Xavier Gómez-Olivé[7,8], Rhian Twine[7], Zané Lombard[4], Scott Hazelhurst[2,9*], Ami S. Bhatt[1,3,10*]┼

[1]Department of Genetics, Stanford University, Stanford, CA, USA
[2]Sydney Brenner Institute for Molecular Bioscience, University of the Witwatersrand, Johannesburg, South Africa
[3]School of Medicine, Stanford University, Stanford, CA, USA
[4]Division of Human Genetics, School of Pathology, Faculty of Health Sciences, National Health Laboratory Service & University of the Witwatersrand, Johannesburg, South Africa
[5]SAMRC Developmental Pathways for Health Research Unit, Department of Paediatrics, University of the Witwatersrand, Johannesburg, South Africa
[6]School of Human Development and Health, University of Southampton, UK
[7]MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa
[8]INDEPTH Network, East Legon, Accra, Ghana
[9]School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg, South Africa
[10]Department of Medicine (Hematology, Blood and Marrow Transplantation), Stanford University, Stanford, CA, USA
*Co-corresponding authors: asbhatt@stanford.edu, Scott.Hazelhurst@wits.ac.za
┼Lead contact

Abstract

While human gut microbiome research often focuses on western populations or nonwestern agriculturalist and hunter-gatherer societies, most of the world's population resides between these extremes. We present the first study evaluating gut microbiome composition in transitioning South African populations using short- and long-read sequencing. We analyzed stool samples from adult females (age 40 - 72) living in rural Bushbuckridge municipality (n=117) or urban Soweto (n=51) and find that these microbiomes are intermediate between those of western industrialized and previously studied non-industrialized African populations. We demonstrate that reference collections are incomplete for nonwestern microbiomes, resulting in within-cohort beta diversity patterns that are in some cases reversed compared to reference-agnostic sequence comparison patterns. To improve reference databases, we generated complete genomes of undescribed taxa, including *Treponema*, *Lentisphaerae*, and *Succinatimonas* species. Our results suggest that South Africa's transitional lifestyle and epidemiological conditions are reflected in gut microbiota compositions, and that these populations contain microbial diversity that remains to be described.

43 # Introduction

44 Comprehensive characterization of the full diversity of the healthy human gut microbiota
45 is essential to contextualize studies of the microbiome in disease. To date, substantial resources
46 have been invested in describing the microbiome of individuals living in the global 'west'
47 (United States, northern and western Europe), including efforts by large consortia such as the
48 Human Microbiome Project (Human Microbiome Project Consortium, 2012) and metaHIT (Qin
49 et al., 2010). Though these projects have yielded valuable descriptions of human gut microbial
50 ecology, they survey only a small portion of the world's citizens at the extreme of industrialized,
51 urbanized lifestyle. It is unclear to what extent these results are generalizable to non-western and
52 non-industrialized populations across the globe.

53 At the other extreme, a relatively smaller number of studies have characterized the gut
54 microbiome composition of non-western individuals practicing traditional lifestyles (Brewster et
55 al., 2019; Gupta et al., 2017), including communities in Venezuela and Malawi (Yatsunenko et
56 al., 2012), hunter-gatherer communities in Tanzania (Fragiadakis et al., 2018; Rampelli et al.,
57 2015; Schnorr et al., 2014; Smits et al., 2017), non-industrialized populations in Tanzania and
58 Botswana (Hansen et al., 2019), and agriculturalists in Peru (Obregon-Tito et al., 2015) and
59 remote Madagascar (Pasolli et al., 2019). However, these cohorts are not representative of how
60 most of the world lives either. Many of the world's communities lead lifestyles between the
61 extremes of an urbanized, industrialized lifestyle and traditional practices. It is a scientific and
62 ethical imperative to include these diverse populations in biomedical research, yet dismayingly
63 many of these intermediate groups are underrepresented or absent from the published
64 microbiome literature.

65 This major gap in our knowledge of the human gut microbiome leaves the biomedical
66 research community ill-poised to relate microbiome composition to human health and disease
67 across the breadth of the world's population. Worldwide, many communities are currently
68 undergoing a transition of diet and lifestyle practice, characterized by increased access to
69 processed foods, diets rich in animal fats and simple carbohydrates, and more sedentary lifestyles
70 (Vangay et al., 2018). This has corresponded with an epidemiological transition in which the
71 burden of disease is shifting from predominantly infectious diseases to include increasing
72 incidence of noncommunicable diseases like obesity and diabetes (Collinson et al., 2014). The
73 microbiome has been implicated in various noncommunicable diseases (Griffiths and
74 Mazmanian, 2018; Helmink et al., 2019; Turnbaugh et al., 2009) and may mediate the efficacy of

75 medical interventions including vaccines (Ciabattini et al., 2019; Hagan et al., 2019), but we
76 cannot evaluate the generalizability of these findings without establishing baseline microbiome
77 characteristics of communities that practice diverse lifestyles and by extension, harbor diverse
78 microbiota. These understudied populations offer a unique opportunity to examine the
79 relationship between lifestyle (including diet), disease, and gut microbiome composition, and to
80 discover novel microbial genomic content.

81 A few previous studies have begun to probe the relationship between lifestyle and
82 microbiome composition in transitional communities (de la Cuesta-Zuluaga et al., 2018; Gupta et
83 al., 2017; Jha et al., 2018; Ou et al., 2013). However, substantial gaps remain in our description
84 of the microbiome in transitional communities. In particular, knowledge of the gut microbiota on
85 the African continent is remarkably sparse. In fact, of 60 studies surveying the gut microbiome in
86 African populations as of mid-2020 (Table S1), 34 (57%) have focused entirely on on children or
87 infants, whose disease risk profile and gut microbiome composition can vary considerably from
88 adults (Lim et al., 2012; Yatsunenko et al., 2012). Additionally, 52 of 60 (87%) of studies of the
89 gut microbiome in Africans employed 16S rRNA gene sequencing or qPCR, techniques which
90 amplify only a tiny portion of the genome and therefore lack genomic resolution to describe
91 species or strains which may share a 16S rRNA sequence but differ in gene content or genome
92 structure. To our knowledge, only five published studies to date have used shotgun
93 metagenomics to describe the gut microbiome of adult populations living in Africa (Campbell et
94 al., 2020; Lokmer et al., 2019; Pasolli et al., 2019; Rampelli et al., 2015; Smits et al., 2017).

95 To address this major knowledge gap, we designed and performed the first research study
96 applying short- and long-read DNA sequencing to study the gut microbiomes of South African
97 individuals for whom 16S rRNA gene sequence data has recently been reported (Oduaran et al.,
98 2020). South Africa is a prime example of a country undergoing rapid lifestyle and
99 epidemiological transition. With the exception of the HIV/AIDS epidemic in the mid-1990s to
100 the mid-2000s, over the past three decades South Africa has experienced a steadily decreasing
101 mortality from infectious disease and an increase in noncommunicable disease (Kabudula et al.,
102 2017a; Santosa and Byass, 2016). Concomitantly, increasingly sedentary lifestyles and changes
103 in dietary habits, including access to calorie-dense processed foods, contribute to a higher
104 prevalence of obesity in many regions of South Africa (Kabudula et al., 2017a), a trend which
105 disproportionately affects women (Ajayi et al., 2016; NCD Risk Factor Collaboration (NCD-
106 RisC) – Africa Working Group, 2017).

107 This study represents the largest shotgun metagenomic dataset of African adults in the
108 published literature to date. In this work, we describe microbial community-scale similarities
109 between urban and rural communities in South Africa, as well as distinct hallmark taxa that
110 distinguish each community. Additionally, we place South Africans in context with microbiome
111 data from other populations globally, revealing the transitional nature of gut microbiome
112 composition in the South African cohorts. We demonstrate that metagenomic assembly of short
113 reads yields novel strain and species draft genomes. Finally, we apply Oxford Nanopore long-
114 read sequencing to samples from the rural cohort and generate complete and near-complete
115 genomes. These include genomes of species that are exclusive to, or more prevalent in,
116 traditional populations, including *Treponema* and *Prevotella* species. As long-read sequencing
117 enables more uniform coverage of AT-rich regions compared to short-read sequencing with
118 transposase-based library preparation, we also generate complete metagenome-assembled AT-
119 rich genomes from less well-described gut microbes including species in the phylum
120 *Melainabacteria*, the class *Mollicutes*, and the genus *Mycoplasma*.
121 Taken together, the results herein offer a more detailed description of gut microbiome
122 composition in understudied transitioning populations, and present complete and contiguous
123 reference genomes that will enable further studies of gut microbiota in nonwestern populations.
124 Importantly, this study was developed with an ethical commitment to engaging both rural and
125 urban community members to ensure that the research was conducted equitably (more details in
126 Supplemental Information). This work underscores the critical need to broaden the scope of
127 human gut microbiome research and include understudied, nonwestern populations to improve
128 the relevance and accuracy of microbiome discoveries to broader populations.

129 # Results

130

131 ***Cohorts and sample collection***

132 We enrolled 190 women aged between 40-72, living in rural villages in the
133 Bushbuckridge Municipality (31.26°E, 24.82°S, n=132) and urban Soweto, Johannesburg
134 (26.25°S, 27.85°E, n=58) and collected a one-time stool sample, as well as point of care blood
135 glucose and blood pressure measurements and a rapid HIV test. Only samples from HIV-
136 negative individuals were analyzed further (n=117 Bushbuckridge, n=51 Soweto). Participants
137 spanned a range of BMI from healthy to overweight; the most common comorbidity reported
138 was hypertension, and many patients reported taking anti-hypertensive medication (18 of 117
139 (15%) in Bushbuckridge, 15 of 51 (29%) in Soweto) (Table 1, Table S2). Additional medications
140 are summarized in Table S2. We extracted DNA from each stool sample and conducted 150 base
141 pair (bp) paired-end sequencing on the Illumina HiSeq 4000 platform. A median of 34.5 million
142 (M) raw reads were generated per sample (range 11.4 M - 100 M), and a median of 11.2 M reads
143 (range 3.2 M - 29.3 M) resulted after pre-processing including de-duplication, trimming, and
144 human read removal (Table S3).

145

146 ***Gut microbial composition***

147 We taxonomically classified sequencing reads against a comprehensive custom reference
148 database containing all microbial genomes in RefSeq and GenBank at scaffold quality or better
149 as of January 2020 (177,626 genomes total). Concordant with observations from 16S rRNA gene
150 sequencing of the same samples (Oduaran et al., 2020), we find that *Prevotella*,
151 *Faecalibacterium*, and *Bacteroides* are the most abundant genera in most individuals across both
152 study sites (Figure 1A, Figure S1, Table S4; species-level classifications in Table S5).
153 Additionally, in many individuals we observe taxa that are uncommon in western microbiomes,
154 including members of the VANISH (Volatile and/or Associated Negatively with Industrialized
155 Societies of Humans) taxa (families *Prevotellaceae*, *Succinovibrionaceae*, *Paraprevotellaceae*,
156 and *Spirochaetaceae*) (Fragiadakis et al., 2018) such as *Prevotella, Treponema*, and
157 *Succinatimonas*, which have been demonstrated to be higher in relative abundance in
158 communities practicing traditional lifestyles compared to westerners (Fragiadakis et al., 2018;
159 Sonnenburg and Sonnenburg, 2019) (Figure 1B, Table S4). The mean relative abundance of each
160 VANISH genus is higher in Bushbuckridge than Soweto, though the difference is not statistically

161 significant for *Prevotella*, *Paraprevotella*, or *Alkalispirochaeta* (Figure 1B, Wilcoxon rank-sum
162 test). Within the Bushbuckridge cohort, we observe a bimodal distribution of the genera
163 *Succinatimonas*, *Succinivibrio*, and *Treponema* (Figure S2A). While we do not identify any
164 participant metadata that associate with this distribution, we observe that VANISH taxa are
165 weakly correlated with one another in metagenomes from both Bushbuckridge and Soweto
166 (Figure S2B-C).
167 Intriguingly, we observed that an increased proportion of reads aligned to the human
168 genome during pre-processing in samples from Soweto compared to Bushbuckridge (Figure S3,
169 Wilcoxon rank sum test $p < 0.0001$). This could potentially indicate higher inflammation and
170 immune cell content or sloughing of intestinal epithelial cells in the urban Soweto cohort
171 compared to rural Bushbuckridge.
172
173 ***Rural and urban microbiomes cluster distinctly in MDS***
174 We hypothesized that lifestyle differences of those residing in rural Bushbuckridge
175 versus urban Soweto might be associated with demonstrable differences in gut microbiome
176 composition. Bushbuckridge and Soweto differ markedly in their population density (53 and
177 6,357 persons per km^2 respectively as of the 2011 census) as well as in lifestyle variables
178 including the prevalence of flush toilets (6.8 vs 91.6% of dwellings) and piped water (11.9 vs
179 55% of dwellings) (additional site demographic information in Table S6) (Statistics South
180 Africa, 2012). Soweto is highly urbanized and has been so for several generations, while
181 Bushbuckridge is classified as a rural community, although it is undergoing rapid
182 epidemiological transition. Bushbuckridge also sees circular rural/urban migrancy typified by
183 some (mostly male) members of a rural community working and living for extended periods in
184 urban areas, while keeping their permanent rural home (Ginsburg et al., 2016). Although our
185 participants all live in Bushbuckridge, this migrancy in the community helps make the boundary
186 between rural and urban lifestyles more fluid. Comparing the two study populations at the
187 community level, we find that samples from the two sites have distinct centroids
188 (PERMANOVA $p < 0.001$, $R^2 = 0.037$) but overlap (Figure 2A), though we note that the
189 dispersion of the Soweto samples is greater than that of the Bushbuckridge samples
190 (PERMDISP2 $p < 0.001$). Across the study population we observe a gradient of Bacteroides and
191 Prevotella relative abundance (Figure S4). This is likely a result of differences in diet across the

192 study population at both sites, as Bacteroides and Prevotella have been proposed as biomarkers
193 of diet and lifestyle (De Filippo et al., 2010; Gorvitovskaia et al., 2016; Yatsunenko et al., 2012).
194 To determine if medication usage was associated with gut microbiome composition, we
195 included each participant's self-reported concomitant medications (summarized in Table S2) to
196 re-visualize the microbiome composition of samples in MDS by class of medication (Figure
197 S5A,B). We find that self-reported medication is not significantly correlated with community
198 composition in this cohort (PERMANOVA $p > 0.05$, Figure S5C) except for in the case of
199 proton pump inhibitors (PPIs) (PERMANOVA $p = 0.026$, $R^2 = 0.0136$). We note that PPIs are
200 one of several drug classes previously found to associate with changes in gut microbiome
201 composition (Maier and Typas, 2017); as only two participants self-report taking PPIs at the time
202 of sampling, additional data is required to evaluate the robustness of this finding in these South
203 African populations.
204
205 ***Rural and urban microbiomes differ in Shannon diversity and species composition***
206 Gut microbiome alpha diversity of individuals living traditional lifestyles has been
207 reported to be higher than those living western lifestyles (De Filippo et al., 2010; Obregon-Tito
208 et al., 2015; Schnorr et al., 2014). In keeping with this general trend, we find that alpha diversity
209 (Shannon) is significantly higher in individuals living in rural Bushbuckridge than urban Soweto
210 (Figure 2B; Wilcoxon rank-sum test, $p < 0.01$). Using DESeq2 to identify microbial genera that
211 are differentially abundant across study sites, we find that genera including *Bacteroides*,
212 *Bifidobacterium,* and *Staphylococcus* are more abundant in individuals living in Soweto (Figure
213 2C, Table S7, species shown in Figure S6). Interestingly, we find microbial genera enriched in
214 gut microbiomes of individuals living in Bushbuckridge that are common to both the
215 environment and the gut, including *Streptomyces* and *Pseudomonas* (Table S7). Typically a soil-
216 associated organism, *Streptomyces* encode a variety of biosynthetic gene clusters and can
217 produce numerous immunomodulatory and anti-inflammatory compounds such as rapamycin
218 and tacrolimus, and it has been suggested that decreased exposure to *Streptomyces* is associated
219 with increased incidence of inflammatory disease and colon cancer in western populations
220 (Bolourian and Mojtahedi, 2018). In addition, we find enrichment of genera in Bushbuckridge
221 that have been previously associated with nonwestern microbiomes including *Succinatimonas*, a
222 relatively poorly-described bacterial genus with only one type species, and Elusimicrobia, a
223 phylum which has been detected in the gut microbiome of rural Malagasy (Pasolli et al., 2019).

224 Additionally, Bushbuckridge samples are enriched for Cyanobacteria as well as Candidatus
225 Melainabacter, a phylum closely related to Cyanobacteria that in limited studies has been
226 described to inhabit the human gut (Di Rienzi et al., 2013; Soo et al., 2014)
227 We find that Bushbuckridge samples have an increased number of bacteriophages (506.1
228 ± 71.7) compared to samples from Soweto (201.5 ± 39.4; p = 8.606e-10). Interestingly, we
229 identify the bacteriophage crAssphage and related crAss-like phages (Guerin et al., 2018), which
230 have recently been described as prevalent constituents of the gut microbiome globally (Edwards
231 et al., 2019), in 32 of 51 participants (63%) in Soweto and 84 of 117 (72%) in Bushbuckridge
232 (difference in prevalence between cohorts not significant, p = 0.28 Fisher's exact test) using 650
233 sequence reads or roughly 1X coverage of the 97 kb genome as a threshold for binary
234 categorization of crAss-like phage presence or absence. Prototypical crAssphage has been
235 hypothesized to infect *Bacteroides* species and a crAss-like phage has been demonstrated to
236 infect *Bacteroides intestinalis*. Though crAss-like phages do not differ between cohorts in terms
237 of prevalence (presence/absence), we observe that both crAss-like phages and *Bacteroides* are
238 enriched in relative abundance in the gut microbiome of individuals living in Soweto compared
239 to Bushbuckridge (Figure 2C).
240
241 ***No strong signals of interaction between human DNA variation and microbiome content***
242 ***detected***
243 We have a very small sample size to assess interaction between human genetic variation
244 and microbiome population. However, as our study is one of the relatively few with both human
245 and microbiome DNA characterized, we performed association tests between key microbiome
246 genera abundance levels and the human DNA. After correcting for multiple testing there were
247 only a few SNPs with borderline statistically significant association with genera abundance
248 levels (Table S8). They occur in genomic regions with no obvious impact on the gut microbiome
249 (see Methods/Supplementary Information). Additionally, we do not observe that samples cluster
250 by self-reported ethnicity of the participant (Figure S7).
251
252 ***South African gut microbiomes share taxa with western and nonwestern populations yet***
253 ***harbor distinct features***
254 To place the microbiome composition of South African individuals in global context with
255 metagenomes from healthy adults living in other parts of the world, we compared publicly

256 available data from four cohorts (Figure 3A, Table S9) comprising adult individuals living in the
257 United States (Human Microbiome Project Consortium, 2012), northern Europe (Sweden)
258 (Bäckhed et al., 2015), rural Madagascar (Pasolli et al., 2019), as well as the Hadza hunter-
259 gatherers of Tanzania (Rampelli et al., 2015). We note the caveat that these samples were
260 collected at different times using different approaches, and that there is variation in DNA
261 extraction, sequencing library preparation and sequencing, all of which may contribute to
262 variation between studies. Recognizing this limitation, we observe that South African samples
263 cluster between western and nonwestern populations[1] in MDS (Figure 3B) as expected, and that
264 the first axis of MDS correlates well with geography and lifestyle (Figure 3C). Additionally, the
265 relative abundance of *Streptomycetaceae*, *Spirochaetaceae*, *Succinivibrionaceae*, and
266 *Bacteroidaceae* are most strongly correlated with the first axis of MDS (Spearman's rho > 0.8):
267 *Bacteroidaceae* decreases with MDS 1 while *Streptomycetaceae*, *Spirochaetaceae*,
268 *Succinivibrionaceae* increase (Figure 3B). These observations suggest that the transitional
269 lifestyle of South African individuals is reflected in their gut microbiome composition. We
270 observe a corresponding pattern of decreasing relative abundance of VANISH taxa across
271 lifestyle and geography (Figure S8).
272 The two South African cohorts also have distinct differences from both nonwestern and
273 western populations, as evidenced by displacement along the second axis of MDS (Figure 3B).
274 To identify the taxa that drive this separation, we analyzed datasets grouped by lifestyle into the
275 general categories of "nonwestern" (Tanzania, Madagascar), "western" (USA, Sweden), and
276 South African (Bushbuckridge and Soweto). We performed statistical analysis using DESeq2 to
277 identify microbial genera that differed significantly in the South African cohort compared to both
278 nonwestern and western categories (with the same directionality of effect in each comparison,
279 e.g. enriched in South Africans compared to both western and nonwestern groups) (Figure S9).
280 We observe that taxa including *Escherichia*, *Lactobacillus*, and *Lactococcus* are lower in relative
281 abundance in South Africans compared to both western and nonwestern categories. Conversely,
282 unclassified bacteria of the phylum Verrucomicrobia are enriched in South Africans.
283 Intriguingly, in this analysis we observe that two crAssphage clades, alpha and delta (Guerin et
284 al., 2018), are lower in abundance in South African participants relative to all other cohorts. This

[1] We use the term "western" to denote western/industrialized populations and "nonwestern" to describe populations not living in the geographic west, as in this case "non-industrialized" does not accurately describe urban Soweto.

285 may suggest a non-uniform geographic distribution of crAssphage clades and/or crAssphage
286 hosts.
287
288 ***Decreased sequence classifiability in nonwestern populations***
289 Given previous observations that gut microbiome alpha diversity is higher in individuals
290 practicing traditional lifestyles (Gupta et al., 2017; Smits et al., 2017; Sonnenburg and
291 Sonnenburg, 2018) and that immigration from a nonwestern nation to the United States is
292 associated with a decrease in gut microbial alpha diversity (Vangay et al., 2018), we
293 hypothesized that alpha diversity would be higher in nonwestern populations including South
294 Africans. We observe that Shannon diversity of the Tanzanian hunter-gatherer cohort is
295 uniformly higher than all other populations (Figure 3D; $p < 0.01$ for all pairwise comparisons;
296 FDR-adjusted Wilcoxon rank sum test) and that alpha diversity is lower in individuals living in
297 the United States compared to all other cohorts (Figure 3D; $p < 0.0001$ for all pairwise
298 comparisons; FDR-adjusted Wilcoxon rank sum test). Surprisingly, we observe comparable
299 Shannon diversity between Madagascar, Bushbuckridge, and Sweden (ns, Wilcoxon rank sum
300 test). However, this could be an artifact of incomplete representation of diverse microbes in
301 existing reference collections.
302 Classification of metagenomic sequences from nonwestern gut microbiomes with
303 existing reference collections is known to be limited (Nayfach et al., 2019; Pasolli et al., 2019),
304 and we observe decreased sequence classifiability in nonwestern populations (Figure 4A).
305 Therefore, we sought orthogonal validation of our observation that South African microbiomes
306 represent a transitional state between traditional and western microbiomes and employed a
307 reference-independent method to evaluate the nucleotide composition of sequence data from
308 each metagenome. We used the sourmash workflow (Brown and Irber, 2016) to compare
309 nucleotide *k*-mer composition of sequencing reads in each sample and ordinated based on
310 angular distance, which accounts for *k*-mer abundance. Using a *k*-mer length of 31 (*k*-mer
311 similarity at k=31 correlates with species-level similarity (Koslicki and Falush, 2016)), we
312 observe clustering reminiscent of the species ordination plot shown in Fig. 3, further supporting
313 the hypothesis that South African microbiomes are transitional (Figure 4B).
314 Previous studies have reported a pattern of higher alpha diversity but lower beta diversity
315 in nonwestern populations compared to western populations (Martínez et al., 2015; Schnorr et
316 al., 2014). Hypothesizing that alpha and beta diversity may be underestimated for populations

317 whose gut microbes are not well-represented in reference collections, we compared beta
318 diversity (distributions of within-cohort pairwise distances) calculated via species Bray-Curtis
319 dissimilarity as well as nucleotide *k*-mer angular distance (Figure 4C-E). Of note, beta diversity
320 is highest in Soweto irrespective of distance measure (Figure 4C). Intriguingly, in some cases we
321 observe that the relationship of distributions of pairwise distance values changes depending on
322 whether species or nucleotide *k*-mers are considered. For instance, considering only species
323 content, Bushbuckridge has less beta diversity than Sweden, but this pattern is reversed when
324 considering nucleotide *k*-mer content (Figure 4D). Further, the same observation is true for the
325 relationship between Madagascar and the United States (Figure 4E). Additionally, we compared
326 species and nucleotide beta diversity within each population using Jaccard distance, which is
327 computed based on shared and distinct features irrespective of abundance. In nucleotide *k*-mer
328 space, all nonwestern populations have greater beta diversity than each western population
329 (Figure S10), though this is not the case when only species are considered. This indicates that gut
330 microbiomes in these nonwestern cohorts have a longer "tail" of lowly abundant organisms
331 which differ between individuals.
332 These observations are critically important to our understanding of beta diversity in the
333 gut microbiome in western and nonwestern communities, as it suggests against the generalization
334 of an inverse relationship between alpha and beta diversity, and in some cases may represent an
335 artifact of limitations in reference databases used for sequence classification.
336
337 ***Improving reference collections via metagenomic assembly***
338 Classification of metagenomic sequencing reads can be improved by assembling
339 sequencing data into metagenomic contigs and grouping these contigs into draft genomes
340 (binning), yielding metagenome-assembled genomes (MAGs). The majority of publications to
341 date have focused on creating MAGs from short-read sequencing data (Almeida et al., 2019;
342 Nayfach et al., 2019; Pasolli et al., 2019), but generation of high-quality MAGs from long-read
343 data from stool samples has been recently reported (Moss et al., 2020). To better characterize the
344 genomes present in our samples, we assembled and binned shotgun sequencing reads from South
345 African samples into MAGs (Figure S11). We generated 3312 MAGs (43 high-quality, 1510
346 medium-quality, and 1944 low-quality) (Bowers et al., 2017) from 168 metagenomic samples,
347 which yielded a set of 1192 non-redundant medium-quality or better representative strain
348 genomes when filtered for completeness greater than 50%, and contamination less than 10% and

349 de-replicated at 99% average nucleotide identity (ANI). This collection of de-replicated genomes
350 includes VANISH taxa including *Prevotella*, *Treponema*, and *Sphaerochaeta* species (Figure
351 S12, Table S10).
352 Interestingly, many MAGs within this set represent organisms that are uncommon in
353 Western microbiomes or not easily culturable, including organisms from the genera *Treponema*
354 and *Vibrio*. As short-read MAGs are typically fragmented and exclude mobile genetic elements,
355 we explored methods to create more contiguous genomes, with a goal of trying to better
356 understand these understudied taxa. We performed long-read sequencing on three samples from
357 participants in Bushbuckridge with an Oxford Nanopore MinION sequencer (taxonomic
358 composition of the three samples shown in Figure S13). Samples were chosen for nanopore
359 sequencing on the basis of molecular weight distribution and total mass of DNA (see Methods).
360 One flow cell per sample generated an average 19.71 Gbp of sequencing with a read N50 of
361 8,275 bp after basecalling. From our three samples, we generated 741 nanopore MAGs
362 (nMAGs), which yielded 35 non-redundant genomes when filtered for completeness greater than
363 50% and contamination less than 10%, and de-replicated at 99% ANI (Table 2, Figure S11,
364 Table S11). All of the de-replicated nMAGs contained at least one full length 16S sequence, and
365 the contig N50 of 28 nMAGs was greater than 1 Mbp (Table S11).
366 We compared assembly statistics between all MAGs and nMAGs, and found that while
367 nMAGs were typically evaluated as less complete by CheckM, the contiguity of nanopore
368 medium- and high-quality MAGs was an order of magnitude higher (mean nMAG N50 of 260.5
369 kb compared to mean N50 of medium- and high-quality MAGs of 15.1 kb) at comparable levels
370 of average coverage (Figure S11, Figure S14). We expect that CheckM under-calculates the
371 completeness of nanopore MAGs due to the homopolymer errors common in nanopore
372 sequencing, which result in frameshift errors when annotating genomes. Indeed, we observe that
373 nanopore MAGs with comparable high assembly size and low contamination to short-read
374 MAGs are evaluated by CheckM as having lower completeness (Figure S14).
375
376 ***Novel genomes generated through nanopore sequencing***
377 When comparing the de-replicated medium- and high-quality nMAGs with the
378 corresponding short-read MAG for the same organism, we find that nMAGs typically include
379 many mobile genetic elements and associated genes that are absent from the short-read MAG,
380 such as transposases, recombinases, phages, and antibiotic resistance genes (Figure 5A).

381 Additionally, a number of the nMAGs are among the first contiguous genomes in their clade. For
382 example, we assembled two single contig, megabase-scale genomes from the genus *Treponema,*
383 a clade that contains various commensal and pathogenic species and is uncommon in the gut
384 microbiota of western individuals (Obregon-Tito et al., 2015; Schnorr et al., 2014). The first of
385 these genomes is a single-contig *Treponema succinifaciens* genome. The type strain of *T.*
386 *succinifaciens*, isolated from the swine gut (Han et al., 2011), is the only genome of this species
387 currently available in public reference collections. Our *T. succinifaciens* genome is the first
388 complete genome of this species from the gut of a human. We assembled a second *Treponema*
389 sp. (Figure S15), which contains an aryl polyene biosynthetic gene cluster and shares 92.1% ANI
390 with *T. succinifaciens*. Additionally, we assembled a 5.08 Mbp genome for *Lentisphaerae sp*.,
391 which has been shown to be significantly enriched in traditional populations (Angelakis et al.,
392 2019). This genome also contains an aryl polyene biosynthetic gene cluster and multiple beta-
393 lactamases, and shares 94% 16S rRNA identity with *Victivallis vadensis*, suggesting a new
394 species or genus of the family *Victivallaceae* and representing the second closed genome for the
395 phylum *Lentisphaerae*.
396 Other nMAGs represent organisms that are prevalent in western individuals but
397 challenging to assemble due to their genome structure. Despite the prevalence of *Bacteroides* in
398 western microbiomes, only three closed *B. vulgatus* genomes are available in RefSeq. We
399 assembled a single contig, 2.68 Mbp *Bacteroides vulgatus* genome that is 65.0% complete and
400 2.7% contaminated and contains at least 16 putative insertion sequences, which may contribute
401 to the lack of contiguous short-read assemblies for this species. Similarly, we assembled a single-
402 contig genome for *Catabacter sp.*, a member of the order *Clostridiales*; the most contiguous
403 *Catabacter* genome in GenBank is in five scaffolded contigs (Parks et al., 2017). The putative
404 *Catabacter sp.* shares 85% ANI with the best match in GenBank, suggesting that it represents a
405 new species within the *Catabacter* genus or a new genus entirely, and it contains a sactipeptide
406 biosynthetic gene cluster. Additionally, we assembled a 3.6 Mbp genome for *Prevotella sp.* (N50
407 = 1.87 Mbp), a highly variable genus that is prevalent in nonwestern microbiomes and associated
408 with a range of effects on host health (Scher et al., 2013). Notably, the first closed genomes of *P.*
409 *copri*, a common species of *Prevotella*, were only recently assembled with nanopore sequencing
410 of metagenomic samples; one from a human stool sample (Moss et al., 2020) and the other from
411 cow rumen (Stewart et al., 2019). *P. copri* had previously evaded closed assembly from short-
412 read sequence data due to the dozens of repetitive insertion sequences within its genome (Moss

413 et al., 2020). Notably, this *Prevotella* assembly contains cephalosporin and beta-lactam
414 resistance genes, as well as an aryl polyene biosynthetic gene cluster.
415 We observed that many long-read assembled genomes were evaluated to be of low
416 completeness despite having contig N50 values greater than 1 Mbp. In investigating this
417 phenomenon, we discovered that many of these genomes had sparse or uneven short-read
418 coverage, leading to gaps in short-read polishing that would otherwise correct small frameshift
419 errors. To polish genomic regions that were not covered with short-reads, we performed long-
420 read polishing on assembled contigs from each sample, and re-binned polished contigs. Long-
421 read polishing improved the completeness of many organisms that are not commonly described
422 in the gut microbiota, due perhaps to their low relative abundance in the average human gut, or
423 to biases in shotgun sequencing library preparation that limit their detection (Figure S16, Figure
424 S17). For example, we generated a 2 Mbp genome that is best classified as a species of the
425 phylum Melainabacteria. Melainabacteria is a non-photosynthetic phylum closely related to
426 Cyanobacteria that has been previously described in the gut microbiome and is associated with
427 consuming a vegetarian diet (Di Rienzi et al., 2013). Melainabacteria have proven difficult to
428 isolate and culture, and the only complete, single-scaffold genome existing in RefSeq was
429 assembled from shotgun sequencing of a human fecal sample (Di Rienzi et al., 2013).
430 Interestingly, our Melainabacteria genome has a GC content of 30.9%, and along with
431 assemblies of a *Mycoplasma sp.* (25.3% GC) and *Mollicutes sp.* (28.1% GC) (Figure S18),
432 represent AT-rich organisms that can be underrepresented in shotgun sequencing data due to the
433 inherent GC bias of transposon insertion and amplification-based sequencing approaches (Sato et
434 al., 2019) (Figure S17). Altogether, these three genomes increased in completeness by an
435 average of 28.5% with long-read polishing to reach an overall average of 70.9% complete. While
436 these genomes meet the accepted standards to be considered medium-quality, it is possible that
437 some or all of these highly contiguous, megabase scale assemblies are complete or near-complete
438 yet underestimated by CheckM due to incomplete polishing.
439 Altogether, we find that *de novo* assembly approaches are capable of generating
440 contiguous, high-quality assemblies for novel organisms, offering potential for investigation into
441 the previously unclassified matter in the microbiomes of these nonwestern communities. In
442 particular, nanopore sequencing was able to produce contiguous genomes for organisms that are
443 difficult to assemble due to repeat structures (*Prevotella sp.*, *Bacteroides vulgatus*), as well as for
444 organisms that exist on the extreme ends of the GC content spectrum (*Mollicutes sp.*,

445 *Melainabacteria sp.*). We observe that long-reads are able to capture a broader range of taxa both
446 at the read and assembly levels when compared to short-read assemblies, and that short- and
447 long-read polishing approaches are able to yield medium-quality or greater draft genomes for
448 these organisms. This illustrates the increased visibility that *de novo* assembly approaches lend to
449 the study of the full array of organisms in the gut microbiome.

450 # Discussion

451 Together with Oduaran *et al.* (Oduaran et al., 2020), we provide the first description of
452 gut microbiome composition in Soweto and Bushbuckridge, South Africa, and to our knowledge,
453 the first effort utilizing shotgun and nanopore sequencing in South Africa to describe the gut
454 microbiome of adults. In doing so, we increase global representation in microbiome research and
455 provide a baseline for future studies of disease association with the microbiome in South African
456 populations, and in other transitional populations.

457 We find that gut microbiome composition differs demonstrably between the
458 Bushbuckridge and Soweto cohorts, further highlighting the importance of studying diverse
459 communities with differing lifestyle practices. Interestingly, even though gut microbiomes of
460 individuals in Bushbuckridge and Soweto share many features and are more similar to each other
461 than to other global cohorts studied, we do observe hallmark taxa associated with westernization
462 are enriched in microbiomes in Soweto. These include *Bacteroides* and *Bifidobacterium*, which
463 have been previously associated with urban communities (Gupta et al., 2017), consistent with
464 Soweto's urban locale in the Johannesburg metropolitan area.

465 We also observe enrichment in relative abundance of crAssphage and crAss-like viruses
466 in Soweto relative to Bushbuckridge, with relatively high prevalence in both cohorts yet lower
467 abundance on average of crAssphage clades alpha and delta compared to several other
468 populations. This furthers recent work which revealed that crAssphage is prevalent across many
469 cohorts globally (Edwards et al., 2019), but found relatively fewer crAssphage sequences on the
470 African continent, presumably due to paucity of available shotgun metagenomic data. Just as
471 shotgun metagenomic sequence data enables the study of viruses, it also enables us to assess the
472 relative abundance of human cells or damaged human cells in the stool. Surprisingly, we observe
473 a high relative abundance of human DNA in the raw sequencing data, which was unexpected.
474 We find a statistically significantly higher relative abundance of human DNA in samples from
475 Soweto compared to those from Bushbuckridge. Future research may help illuminate the
476 potential reason for this finding, which may include a higher proportion of epithelium disrupting,
477 invasive bacteria or parasites in Soweto vs. Bushbuckridge, and in South Africa, in general,
478 compared to other geographic settings. Alternatively, this may also be attributable to a higher
479 baseline of intestinal inflammation and fecal shedding of leukocytes. Without additional
480 information, it is difficult to speculate as to the reason for this finding.

481 We find that individuals in Bushbuckridge are enriched in VANISH taxa including
482 *Succinatimonas*, which has been recently reported to associate with microbiomes from
483 individuals practicing traditional lifestyles (Pasolli et al., 2019). Intriguingly, several VANISH
484 taxa (*Succinatimonas*, *Succinivibrio*, *Treponema*) display bimodal distributions in the
485 Bushbuckridge cohort. We hypothesize that this bimodality could be caused by differences in
486 lifestyle and/or environmental factors including diet, history of hospitalization or exposure to
487 medicines, physical properties of the household dwelling, differential treatment of drinking water
488 across the villages comprising Bushbuckridge. Additionally this pattern may be explained by
489 participation in migration to and from urban centers (or sharing a household with a migratory
490 worker). A higher proportion of men in the community engage in this pattern of rural-urban
491 migration (Ginsburg et al., 2016), but it is possible that sharing a household with a cyclical
492 worker could influence gut microbiome composition via horizontal transmission (Brito et al.,
493 2019).
494 Despite the fact that host genetics explain relatively little of the variation in microbiome
495 composition (Rothschild et al., 2018), we do observe a small number of taxa that associate with
496 host genetics in this population. Future work is required for replication and to determine whether
497 these organisms are interacting with the host and whether they are associated with host health.
498 Additionally, we demonstrate marked differences between South African cohorts and
499 other previously studied populations living on the African continent and western countries.
500 Broadly, we find that South African microbiomes reflect the transitional nature of their
501 communities in that they overlap with western and nonwestern populations. Tremendous human
502 genetic diversity exists within Africa, and our work reveals that there is a great deal of as yet
503 unexplored microbiome diversity as well. In fact, we find that microbiome beta diversity within
504 communities may be systematically underestimated by incomplete reference databases: taxa that
505 are unique to individuals in nonwestern populations are not present in reference databases and
506 therefore not included in beta diversity calculations. Though it has been reported that nonwestern
507 and traditional populations tend to have higher alpha diversity but lower beta diversity compared
508 to western populations, we show that this pattern is not universally upheld when reference-
509 agnostic nucleotide comparisons are performed. By extension, we speculate that previous claims
510 that beta diversity inversely correlates with alpha diversity may have been fundamentally limited
511 by study design in some cases. Specifically, the disparity between comparing small, homogenous
512 African populations with large, heterogenous western ones constitutes a significant statistical

513 confounder, potentially preventing a valid assessment of beta diversity between groups.
514 Furthermore, alpha and beta diversity comparisons based on species-level taxonomic assignment
515 may be further confounded due to the presence of polyphyletic clades in organisms like
516 *Prevotella copri* (Parks et al., 2020; Tett et al., 2019) which are highly abundant in gut
517 microbiomes of nonwestern individuals.

518 Through a combination of short-read and long-read sequencing, we successfully
519 assembled contiguous, complete genomes for many organisms that are underrepresented in
520 reference databases, including genomes that are commonly considered to be enriched in or
521 limited to populations with traditional lifestyles including members of the VANISH taxa (e.g.,
522 *Treponema sp., Treponema succinifaciens*). The phylum *Spirochaetes*, namely its constituent
523 genus *Treponema*, is considered to be a marker of traditional microbiomes and has not been
524 detected in high abundance in human microbiomes outside of those communities (Angelakis et
525 al., 2019; Obregon-Tito et al., 2015). Here, we identify *Spirochaetes* in the gut microbiome of
526 individuals in urban Soweto, demonstrating that this taxon is not exclusive to traditional, rural
527 populations, though we observe that relative abundance is higher on average in traditional
528 populations. Generation of additional genomes of VANISH taxa and incorporation of these
529 genomes into reference databases will allow for increased sensitivity to detect these organisms in
530 metagenomic data. Additionally, these genomes facilitate comparative genomics of understudied
531 gut microbes and allow for functional annotation of potentially biologically relevant functional
532 pathways. We note that many of these genomes (e.g., Melainabacteria, *Succinatimonas*) are
533 enriched in the gut microbiota of Bushbuckridge participants relative to Soweto, highlighting the
534 impact of metagenomic assembly to better resolve genomes present in rural populations.

535 We produced genomes for organisms that exist on the extremes of the GC content
536 spectrum, such as *Mycoplasma sp.*, *Mollicutes sp.*, and *Melainabacteria sp*. We find that these
537 organisms are sparsely covered by short-read sequencing, illustrating the increased range of non-
538 amplification based sequencing approaches, such as nanopore sequencing. Interestingly, these
539 assemblies are evaluated as only medium-quality by CheckM despite having low measurements
540 of contamination, as well as genome lengths and gene counts comparable to reference genomes
541 from the same phylogenetic clade. We hypothesize that sparse short-read coverage leads to
542 incomplete polishing and therefore retention of small frameshift errors, which are a known
543 limitation of nanopore sequencing (Tyler et al., 2018). Further evaluation of 16S or long-read

544 sequencing of traditional and western populations can identify whether these organisms are
545 specific to certain lifestyles, or more prevalent but poorly detected with shotgun sequencing.
546 While we find that the gut microbiome composition of the two South African cohorts
547 described herein reflects their lifestyle transition, we acknowledge that these cohorts are not
548 necessarily representative of all transitional communities in South Africa or other parts of the
549 world which differ in lifestyle, diet, and resource access. Hence, further work remains to describe
550 the gut microbiota in detail of other such understudied populations. This includes a detailed
551 characterization of parasites present in microbiome sequence data, an analysis that we did not
552 undertake in this study but would be of great interest. These organisms have been detected in the
553 majority of household toilets in nearby KwaZulu-Natal province (Trönnberg et al., 2010), and
554 may interact with and influence microbiota composition (Leung et al., 2018).
555 Our study has several limitations. Although the publicly available sequence data from
556 other global cohorts were generated with similar methodology to our study, it is possible that
557 batch effects exist between datasets generated in different laboratories that may explain some
558 percentage of the global variation we observe. Additionally, while nanopore sequencing is able
559 to broaden our range of investigation, we illustrate that our ability to produce well-polished
560 genomes at GC content extremes is limited. This may affect our ability to accurately call gene
561 lengths and structures, although iterative long-read polishing improves our confidence in these
562 assemblies. Future investigation of these communities using less biased, higher coverage short-
563 read approaches or more accurate long-read sequencing approaches, such as PacBio circular
564 consensus sequencing, may improve assembly qualities. Additionally, long-read sequencing of
565 samples from a wider range of populations can identify whether the genomes identified herein
566 are limited to traditional and transitional populations, or more widespread. Further, future
567 improvements in error rate of long-read sequencing may obviate the need for short-read
568 polishing altogether.
569 Taken together, our results emphasize the importance of generating sequence data from
570 diverse transitional populations to contextualize studies of health and disease in these
571 individuals. To do so with maximum sensitivity and precision, reference genomes must be
572 generated to classify sequencing reads from these metagenomes. Herein, we demonstrate the
573 discrepancies in microbiome sequence classifiability across global populations and highlight the
574 need for more comprehensive reference collections. Recent efforts have made tremendous
575 progress in improving the ability to classify microbiome data through creating new genomes via

576 metagenomic assembly (Almeida et al., 2019; Nayfach et al., 2019; Pasolli et al., 2019), and here
577 we demonstrate the application of short- and long-read metagenomic assembly techniques to
578 create additional genome references. Our application of long-read sequencing technology to
579 samples from South African individuals has demonstrated the ability to generate highly
580 contiguous MAGs and shows immense potential to expand our reference collections and better
581 describe microbiomes throughout diverse populations globally. In the future, microbiome studies
582 may utilize a combination of short- and long-read sequencing to maximize information output,
583 perhaps performing targeted Nanopore sequencing of samples that are likely to contain the most
584 novelty on the basis of short-read data.

585 The present study was conducted in close collaboration between site staff and researchers
586 in Bushbuckridge and Soweto as well as microbiome experts both in South Africa and the United
587 States, and community member feedback was considered at multiple phases in the planning and
588 execution of the study (see Oduaran *et al.* 2020 for more information). Tremendous research
589 efforts have produced detailed demographic and health characterization of individuals living in
590 both Bushbuckridge and Soweto (Kabudula et al., 2017a, 2017b; Ramsay et al., 2016; Richter et
591 al., 2007) and it is our hope that microbiome data can be incorporated into this knowledge
592 framework in future studies to uncover disease biomarkers or microbial associations with other
593 health and lifestyle outcomes. More broadly, we feel that this is an example of a framework for
594 conducting microbiome studies in an equitable manner, and we envision a system in which
595 future studies of microbiome composition can be carried out to achieve detailed characterization
596 of microbiomes globally while maximizing benefit to all participants and researchers involved.

597 # Methods

598 ## Cohort selection

599 Stool samples were collected from women aged 40-72 years in Soweto, South Africa and
600 Bushbuckridge Municipality, South Africa. Participants were recruited on the basis of
601 participation in AWI-Gen (Ramsay et al., 2016), a previous study in which genotype and
602 extensive health and lifestyle survey data were collected. Human subjects research approval was
603 obtained (Stanford IRB 43069, University of the Witwatersrand Human Research Ethics
604 Committee M160121, Mpumalanga Provincial Health Research Committee MP_2017RP22_851)
605 and informed consent was obtained from participants for all samples collected. Stool samples
606 were collected and preserved in OmniGene Gut OMR-200 collection kits (DNA Genotek).
607 Samples were frozen within 60 days of collection as per manufacturer's instructions, followed by
608 long-term storage at -80°C. As the enrollment criteria for our study included previous
609 participation in a larger human genomics project (Ramsay et al., 2016), we had access to self-
610 reported ethnicity for each participant (BaPedi, Ndebele, Sotho, Tsonga, Tswana, Venda, Xhosa,
611 Zulu, Other, or Unknown). Samples from participants who tested HIV-positive or who did not
612 consent to an HIV test were not analyzed.

613 ## Metagenomic sequencing of stool samples

614 DNA was extracted from stool samples using the QIAamp PowerFecal DNA Kit
615 (QIAGEN) according to the manufacturer's instructions except for the lysis step, in which
616 samples were lysed using the TissueLyser LT (QIAGEN) (30 second oscillations/3 minutes at
617 30Hz). DNA concentration of all DNA samples was measured using Qubit Fluorometric
618 Quantitation (DS DNA High-Sensitivity Kit, Life Technologies). DNA sequencing libraries were
619 prepared using the Nextera XT DNA Library Prep Kit (Illumina). Final library concentration was
620 measured using Qubit Fluorometric Quantitation and library size distributions were analyzed
621 with the Bioanalyzer 2100 (Agilent). Libraries were multiplexed and 150 base pair paired-end
622 reads were generated on the HiSeq 4000 platform (Illumina). Samples with greater than
623 approximately 300 ng remaining mass and a peak fragment length of greater than 19,000 bp
624 (with minimal mass under 4,000 bp) as determined by a TapeStation 2200 (Agilent
625 Technologies, Santa Clara, CA) were selected for nanopore sequencing. Nanopore sequencing
626 libraries were prepared using the 1D Genomic DNA by Ligation protocol (ONT, Oxford UK)

627 following standard instructions. Each library was sequenced with a full FLO-MIN106D R9
628 Version Rev D flow cell on a MinION sequencer for at least 60 hours.

629 Computational methods

630 R code for analysis and figure generation will be made available on Github upon publication.
631
632 *Preprocessing*
633 Stool metagenomic sequencing reads were trimmed using TrimGalore v0.5.0 (Krueger), a
634 wrapper for CutAdapt v1.18 (Martin, 2011), with a minimum quality score of 30 for trimming (--
635 q 30) and minimum read length of 60 (--length 60). Trimmed reads were deduplicated to remove
636 PCR and optical duplicates using seqtk rmdup v1.3-r106 with default parameters. Reads aligning
637 to the human genome (hg19) were removed using BWA v0.7.17-r1188 (Li and Durbin, 2009).
638 To assess the microbial composition of our short-read sequencing samples, we used the Kraken
639 v2.0.8-beta taxonomic sequence classifier with default parameters (Wood and Salzberg, 2014)
640 and a comprehensive custom reference database containing all bacterial and archaeal genomes in
641 GenBank assembled to "complete genome," "chromosome," or "scaffold" quality as of January
642 2020. Bracken v2.0.0 was then used to re-estimate abundance at each taxonomic rank (Lu et al.,
643 2017).
644
645 *Additional data*
646 Published data from additional populations were downloaded via the NCBI Sequence
647 Read Archive (SRA) or European Nucleotide Archive (Table S9) and preprocessed and
648 taxonomically classified as described above. For datasets containing longitudinal samples from
649 the same individual, one unique sample per individual was chosen (the first sample from each
650 individual was chosen from the United States Human Microbiome Project cohort).
651
652 *K-mer sketches*
653 *K*-mer sketches were computed using sourmash v2.0.0 (Brown and Irber, 2016). Low
654 abundance *k*-mers were trimmed using the "trim-low-abund.py" script from the khmer package
655 (Crusoe et al., 2015) with a *k*-mer abundance cutoff of 3 (-C 3) and trimming coverage of 18 (-Z
656 18). Signatures were computed for each sample using the command "sourmash compute" with a
657 compression ratio of 1000 (--scaled 1000) and *k*-mer lengths of 21, 31, and 51 (-k 21,31,51).
658 Two signatures were computed for each sample - one signature tracking *k*-mer abundance (--

track-abundance flag) for angular distance comparisons, and one without this flag for Jaccard distance comparisons. Signatures at each length of *k* were compared using "sourmash compare" with default parameters and the correct length of *k* specified with the -k flag.

Statistical analysis and plotting

Statistical analyses were performed using R v4.0.0 (R Core Team, 2019) with packages MASS v7.3-51.5 (Venables and Ripley, 2002), stats (R Core Team, 2019), ggsignif v0.6.0 (Ahlmann-Eltze, 2019), and ggpubr v0.2.5 (Kassambara, 2020). Alpha and beta diversity were calculated using the vegan package v2.5-6 (Oksanen et al., 2019). Wilcoxon rank-sum tests were used to compare alpha and beta diversity between cohorts. Count data were normalized via cumulative sum scaling and log2 transformation (Paulson et al., 2013) prior to MDS. Data separation in MDS was assessed via PERMANOVA (permutation test with pseudo F ratios) using the adonis function from the vegan package. Differential microbial features between individuals living in Soweto and Bushbuckridge were identified from unnormalized count data output from kraken2 classification and bracken abundance re-estimation and filtered for 20% prevalence and at least 1000 sequencing reads using DESeq2 (Love et al., 2014). Plots were generated in R using the following packages: cowplot v1.0.0 (Wilke, 2019), DESeq2 v1.24.0 (Love et al., 2014), dplyr v0.8.5 (Wickham et al., 2020), genefilter v1.66.0 (Gentleman et al., 2019), ggplot2 v3.3.0 (Wickham, 2016), ggpubr v0.2.5, ggrepel v0.8.2 (Slowikowski, 2020), ggsignif v0.6.0, gtools v3.8.2 (Warnes et al., 2020), harrietr v0.2.3 (Gonçalves da Silva, 2017), MASS v7.3-51.5, reshape2 v1.4.3 (Wickham, 2007), and vegan v2.5-6.

Genome assembly, binning, and evaluation

Short-read metagenomic data were assembled with MEGAHIT v1.1.3 (Li et al., 2016) and binned into draft genomes as previously described (Bishara et al., 2018). Briefly, short reads were aligned to assembled contigs with BWA v0.7.17 (Li and Durbin, 2009) and contigs were subsequently binned into draft genomes with MetaBAT v2:2.13 (Kang et al., 2015). Bins were evaluated for size, contiguity, completeness, and contamination with QUAST v5.0.0 (Gurevich et al., 2013), CheckM v1.0.13 (Parks et al., 2015), Prokka v1.13 (Seemann, 2014), Aragorn v1.2.38 (Laslett and Canback, 2004), and Barrnap v0.9 (https://github.com/tseemann/barrnap/). We referred to published guidelines to designate genome quality (Bowers et al., 2017). Individual contigs from all assemblies were assigned taxonomic classifications with Kraken

691 v2.0.8 (Bowers et al., 2017; Wood and Salzberg, 2014). Genome sets were filtered for
692 completeness greater than 50% and contamination less than 10% as evaluated by CheckM, and
693 de-replicated using dRep v2.5.4 (Olm et al., 2017) with ANI threshold to form secondary clusters
694 (-sa) at 0.99 (strain-level) or 0.95 (species-level).

695 Long-read data were assembled with Lathe (Moss et al., 2020) as previously described.
696 Briefly, Lathe implements basecalling with Guppy v2.3.5, assembly with Flye v2.4.2 (Lin et al.,
697 2016), short-read polishing with Pilon v1.23 (Walker et al., 2014), and circularization with
698 Circlator (Hunt et al., 2015) and Encircle (Moss et al., 2020). Binning, classification, and de-
699 replication were performed as described above. Additional long-read polishing was performed
700 using four iterations of polishing with Racon v1.4.10 (Vaser et al., 2017) and long-read
701 alignment using minimap2 v2.17-r941 (Li, 2018), followed by one round of polishing with
702 Medaka v0.11.5 (https://github.com/nanoporetech/medaka).

703 Direct comparisons between nMAGs and corresponding MAGs were performed by de-
704 replicating high- and medium-quality nMAGs with MAGs assembled from the same sample.
705 MAGs sharing at least 99% ANI with an nMAG were aligned to the nMAG regions using
706 nucmer v3.1 and uncovered regions of the nMAG were annotated with prokka 1.14.6,
707 VIBRANT v1.2.1, and ResFams v1.2. Taxonomic trees were plotted with Graphlan v1.1.3
708 (Asnicar et al., 2015).

709 To construct phylogenetic trees, reference 16S sequences were downloaded from the
710 Ribosomal Database Project (Release 11, update 5, September 30, 2016) (Cole et al., 2014) and
711 16S sequences were identified from nanopore genome assemblies using Barrnap v0.9
712 (https://github.com/tseemann/barrnap/). Sequences were aligned with MUSCLE v3.8.1551
713 (Edgar, 2004) with default parameters. Maximum-likelihood phylogenetic trees were constructed
714 from the alignments with FastTree v2.1.10 (Edgar, 2004; Price et al., 2010) with default settings
715 (Jukes-Cantor + CAT model). Support values for branch splits were calculated using the
716 Shimodaira-Hasegawa test with 1,000 resamples (default). Trees were visualized with FigTree
717 v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).

718 Data availability

719 All shotgun sequence data generated by this study, as well as metagenome-assembled
720 genome sequences, will be deposited in a publicly available reference database (NCBI Sequence
721 Read Archive or European Nucleotide Archive) and released upon publication.

722 Participant-level metadata (age, BMI, blood pressure measurements, and concomitant
723 medications) and human genetic data will be deposited in the European Genome-phenome
724 Archive and released upon publication.

725 Acknowledgements

726 We thank the participants in our study for taking part in this research. Additionally, we
727 thank the Bushbuckridge Community Advisory Group for their thoughtful recommendations on
728 study procedure. We thank Karen Andrade for her contributions in planning the 2019
729 Community Advisory Group workshop. We thank the INDEPTH consortium for their support of
730 this project. We thank the numerous fieldworkers at the Soweto DPHRU and Agincourt HDSS
731 who enrolled participants and collected data. In particular, we thank Melody Mabuza, the field
732 worker at Agincourt HDSS who oversaw collection of Bushbuckridge participant enrollment,
733 and Jackson Mabasa, who managed sample collection in Soweto. We thank Michèle Ramsay
734 (AWI-Gen PI), Yusuf Ismail, and Amanda Haye for their contributions to organizational and
735 sample processing aspects of the project. We thank the Stanford Research Computing Center and
736 Ben Siranosian for their contributions to computational infrastructure and support.

737 Funding

738 This work was supported in part by a grant from the Stanford Center for Innovation in
739 Global Health and by NIH grant P30 CA124435, which supports the following Stanford Cancer
740 Institute Shared Resource: the Genetics Bioinformatics Service Center. A.S.B was supported by
741 the Rosenkranz prize and by an R01 AI148623 from the National Institute of Allergy and
742 Infectious Diseases. F.B.T. was supported by the National Science Foundation Graduate
743 Research Fellowship and the Stanford Computational, Evolutionary, and Human Genetics Pre-
744 Doctoral Fellowship. D.G.M was supported by the Stanford Graduate Fellowships in Science
745 and Engineering program. O.H.O. was partially supported by a Fogarty Global Health Equity
746 Scholar award (TW009338). ANW is supported by the Fogarty International Centre, National
747 Institutes of Health under award number K43TW010698. The work was further supported by the
748 South African National Research Foundation (CPRR160421162721) and a seed grant from the
749 African Partnership for Disease Control. The AWI-Gen project is supported by the National
750 Human Genome Research Institute (U54HG006938) as part of the H3A Consortium. The
751 MRC/Wits Rural Public Health and Health Transitions Research Unit and Agincourt Health and

752 Socio-Demographic Surveillance System, a node of the South African Population Research
753 Infrastructure Network (SAPRIN), is supported by the Department of Science and Innovation,
754 the University of the Witwatersrand and the Medical Research Council, South Africa, and
755 previously the Wellcome Trust, UK (grants 058893/Z/99/A; 069683/Z/02/Z; 085477/Z/08/Z;
756 085477/B/08/Z).This paper describes the views of the authors and does not necessarily represent
757 the official views of the National Institutes of Health (USA).

Main Tables

Table 1. Participant characteristics

	Site	Mean	Standard deviation	Range
Age	**Bushbuckridge**	55.52	7.79	43 - 72
	Soweto	54.1	5.86	43 - 64
BMI	**Bushbuckridge**	32.35	8.00	21.2 - 59*
	Soweto	36.05	9.25	20.42 - 58.62
Systolic blood pressure	**Bushbuckridge**	137	18.28	101 - 189
	Soweto	134	22.54	96 - 193
Diastolic blood pressure	**Bushbuckridge**	84	12.12	54 - 119
	Soweto	90	14.37	58 - 119

*One participant's BMI measurement was excluded on the basis of the recorded value being too low to be physiologically possible and deemed to have been recorded in error. We could not validate the correct BMI for this participant and thus have omitted them from the BMI summary statistics.

763 **Table 2. Medium- and high-quality genomes assembled from nanopore sequencing**

Classification	Size (Mb)	% GC	N50 (Mb)	Quality	16S	Antibiotic Resistance Genes	Phages	Transposases	Biosynthetic Gene Clusters	Polishing
Alistipes putredinis	1.91	53.1	1.91	Medium	2	1	1	1	0	Short
Anaerotruncus sp.	2.04	43.7	2.04	Medium	2	2	2	4	1	Short
Bacilli bacterium	1.46	26.2	1.46	Medium	1	0	2	1	1	Short
Bacteroidales bacterium	2.67	47.3	1.80	High	3	0	4	16	0	Short
Bacteroidales bacterium	2.79	49.8	2.79	High	4	3	0	29	0	Short
Bacteroidales bacterium	1.7	56.6	1.70	Medium	1	1	0	6	0	Short
Bacteroides sp.	2	48.2	1.59	High	3	1	0	7	0	Short
Bacteroides sp.	2.82	43.3	2.00	Medium	6	1	3	31	0	Short
Bacteroides vulgatus	2.68	42.7	2.68	Medium	3	0	0	14	0	Short
Candidatus Melainabacteria	2	30.9	2.00	Medium	1	0	4	0	0	Long and Short
Catabacter sp.	1.65	46.4	1.65	Medium	1	2	1	0	1	Long and Short
Clostridiales bacterium	2.03	57.9	0.60	Medium	4	2	2	6	1	Short
Clostridiales bacterium	1.53	47.3	1.53	Medium	1	1	1	1	1	Short
Clostridiales bacterium	1.95	49.6	0.73	Medium	3	5	2	1	1	Short
Clostridiales bacterium	2.24	48.7	0.58	Medium	2	3	3	12	1	Short
Clostridiales bacterium	2.65	42.8	2.65	Medium	3	0	3	6	2	Short
Clostridiales bacterium	1.32	45.2	0.79	Medium	1	3	2	4	1	Short
Clostridiales bacterium	1.61	46.9	1.61	Medium	1	1	2	0	0	Short
Clostridium sp.	1.53	25.2	1.53	Medium	1	0	2	1	0	Short
Clostridium sp.	1.3	46.9	1.30	Medium	1	2	1	0	0	Short
Clostridium sp.	2.01	28.8	2.01	Medium	3	2	3	3	0	Short
Clostridium sp.	1.14	29.1	1.14	Medium	1	0	1	0	0	Short
Clostridium sp.	2.44	52.5	2.23	High	3	6	3	1	3	Short
Eubacterium	2	44.5	0.63	Medium	2	1	1	5	0	Short
Lachnospiraceae bacterium	3.38	43.6	1.94	Medium	4	7	2	10	0	Short
Lachnospiraceae bacterium	3.81	43.6	2.83	Medium	4	6	2	28	2	Short
Lentisphaeria bacterium	5.08	57.5	5.08	Medium	3	3	4	84	1	Long and Short
Mollicutes bacterium	1.68	28.1	1.49	Medium	2	1	1	2	0	Long and Short
Mycoplasma sp.	1.17	25.3	1.12	Medium	2	2	0	1	0	Long and Short
Oscillibacter sp.	1.13	57.4	0.17	Medium	1	0	2	2	0	Short
Porphyromonadaceae bacterium	2.97	47.4	2.97	Medium	5	1	1	9	0	Short
Prevotella sp.	3.29	43.6	1.14	Medium	6	3	2	17	1	Long and Short
Ruminococcaceae bacterium	1.95	38.4	0.80	Medium	4	0	1	8	0	Short
Ruminococcaceae bacterium	2.27	51.4	2.27	High	3	4	2	4	1	Short
Ruminococcaceae bacterium	1.78	58.3	1.78	Medium	3	3	0	9	0	Short
Treponema sp.	2.06	41.6	2.06	Medium	3	0	2	2	1	Short
Treponema succinifaciens	2.55	39.1	2.55	High	4	0	0	15	0	Short
uncultured Ruminococcus	1.59	44.0	1.34	Medium	2	2	0	2	1	Short
uncultured Ruminococcus	2.08	46.9	2.08	Medium	5	2	6	8	1	Short

Figures

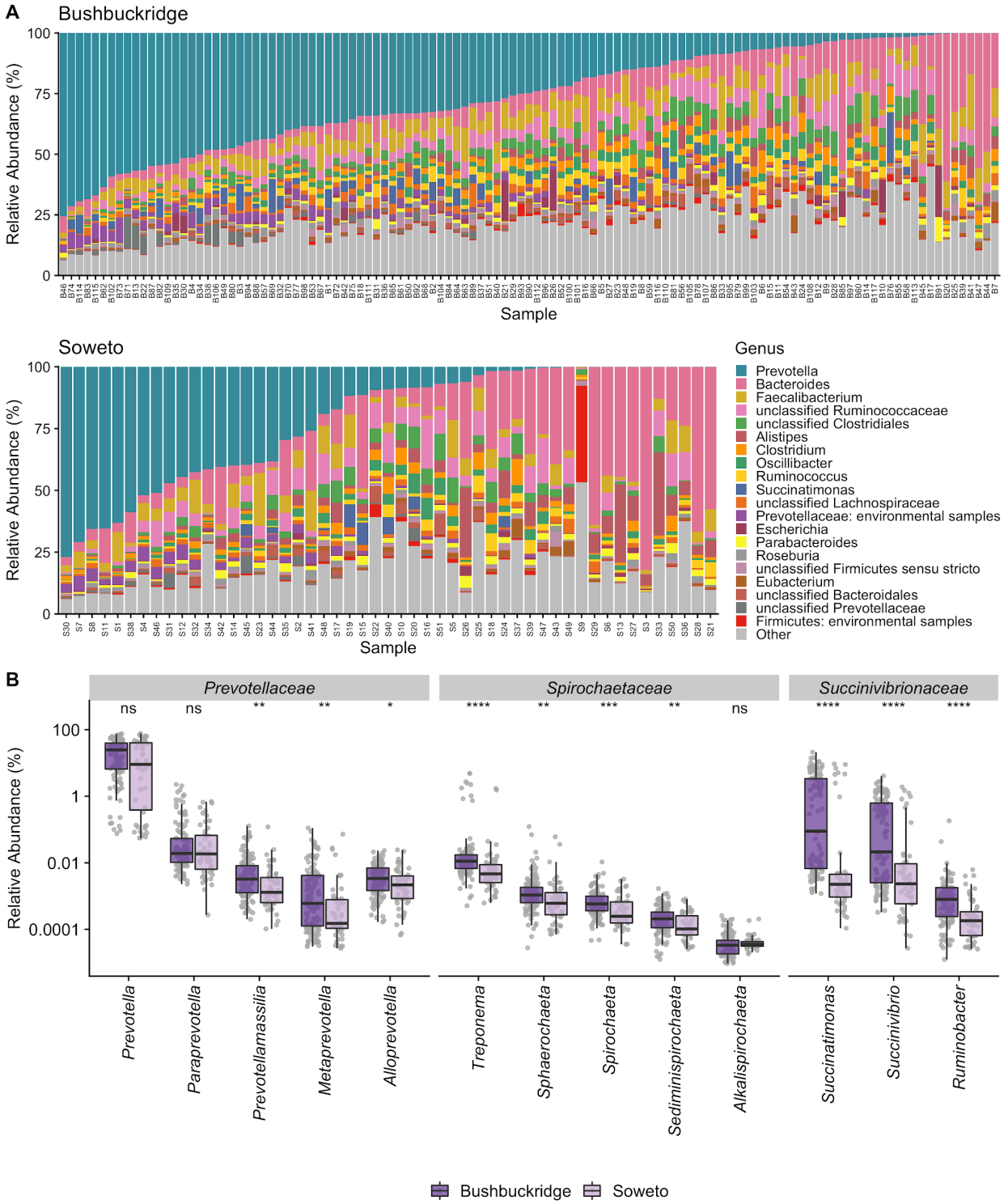


Figure 1. Taxonomic composition of South African study participants

Sequence data were taxonomically classified using Kraken2 with a database containing all genomes in GenBank of scaffold quality or better as of January 2020.

769 (A) Top 20 genera by relative abundance for samples from participants in Bushbuckridge and Soweto, sorted by
770 decreasing *Prevotella* abundance. *Prevotella*, *Faecalibacterium*, and *Bacteroides* are the most prevalent genera
771 across both study sites.
772 (B) Relative abundance of VANISH genera by study site, grouped by family. A pseudocount of 1 read was added to
773 each sample prior to relative abundance normalization in order to plot on a log scale, as the abundance of some
774 genera in some samples is zero. Relative abundance values of most VANISH genera are higher on average in
775 participants from Bushbuckridge than Soweto (Wilcoxon rank-sum test, significance values denoted as follows: (*)
776 $p < 0.05$, (**) $p < 0.01$, (***) $p < 0.001$, (****) $p < 0.0001$, (ns) not significant). Upper and lower box plot whiskers
777 represent the highest and lowest values within 1.5 times the interquartile range, respectively.

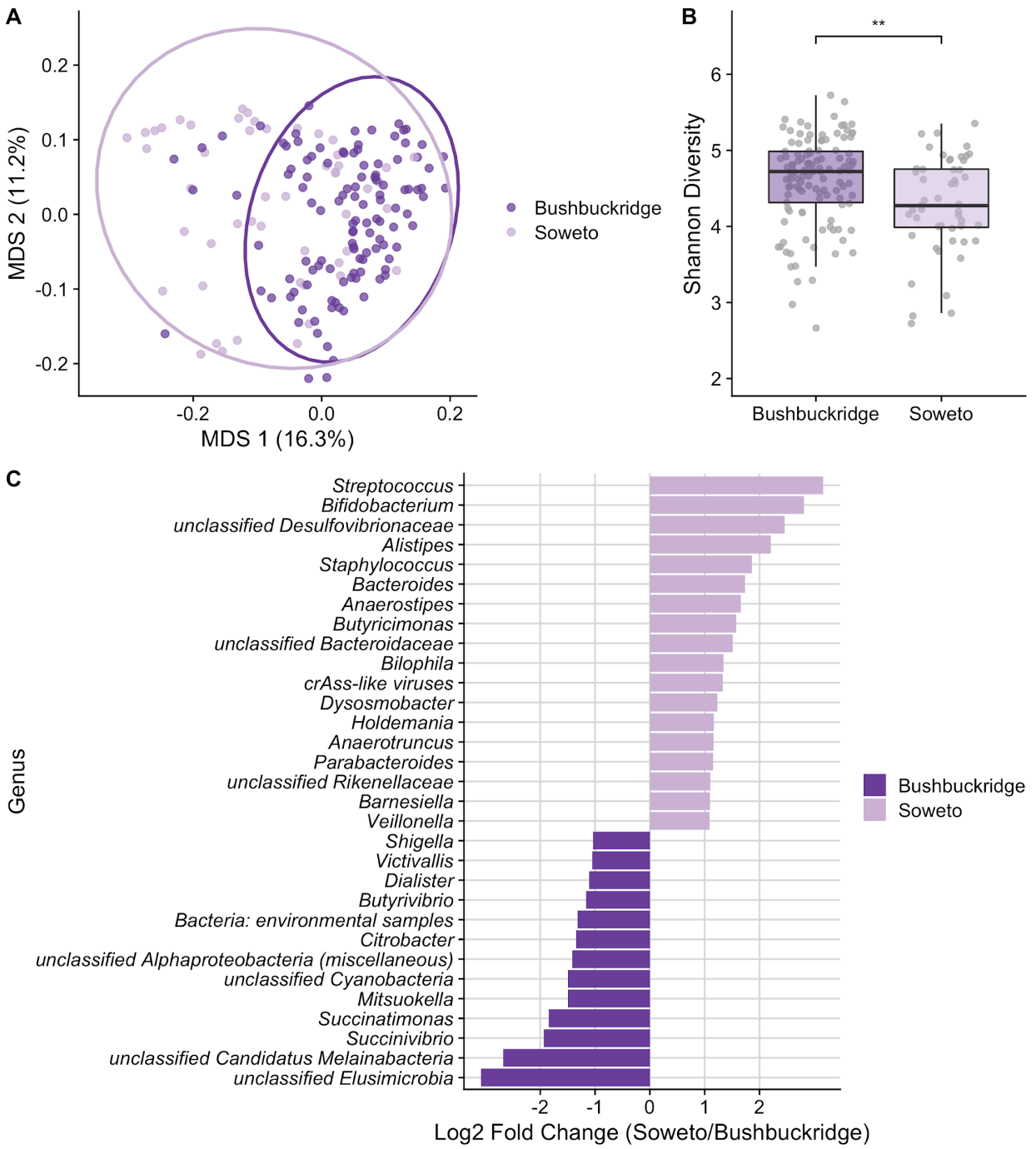


Figure 2. Comparison of Bushbuckridge and Soweto microbiomes

(A) Multidimensional scaling of pairwise Bray-Curtis distance between samples (CSS-normalized counts). Samples from Soweto have greater dispersion than samples from Bushbuckridge (PERMDISP2 $p < 0.001$).

(B) Shannon diversity calculated on species-level taxonomic classifications for each sample. Samples from Bushbuckridge are higher in alpha diversity than samples from Soweto (Wilcoxon rank-sum test, $p < 0.001$). Upper and lower box plot whiskers represent the highest and lowest values within 1.5 times the interquartile range, respectively.

786 (C) DESeq2 identifies microbial genera that are differentially abundant in rural Bushbuckridge compared to the
787 urban Soweto cohort. Features with log2 fold change greater than one are plotted (full results in Table S7).
788

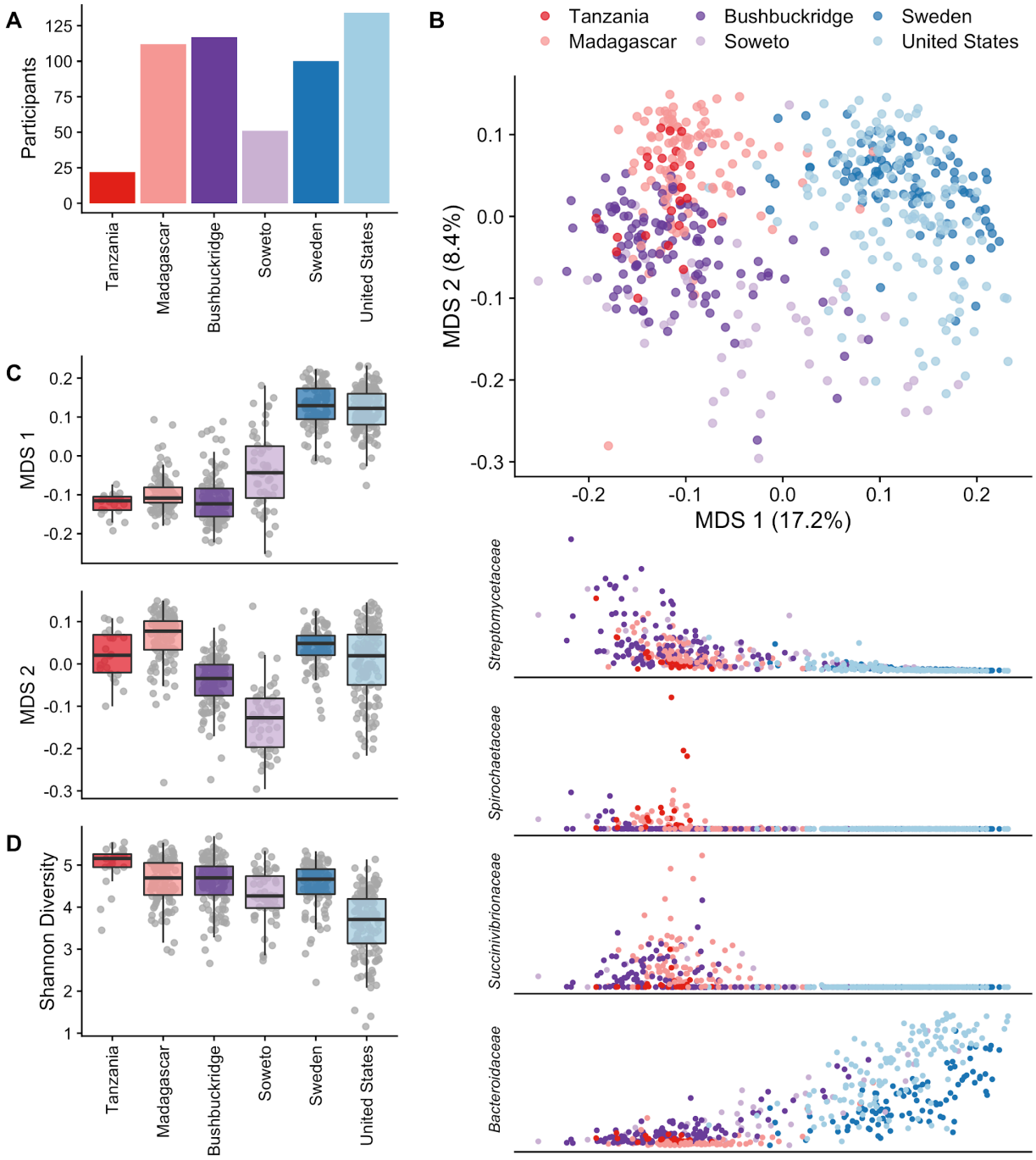


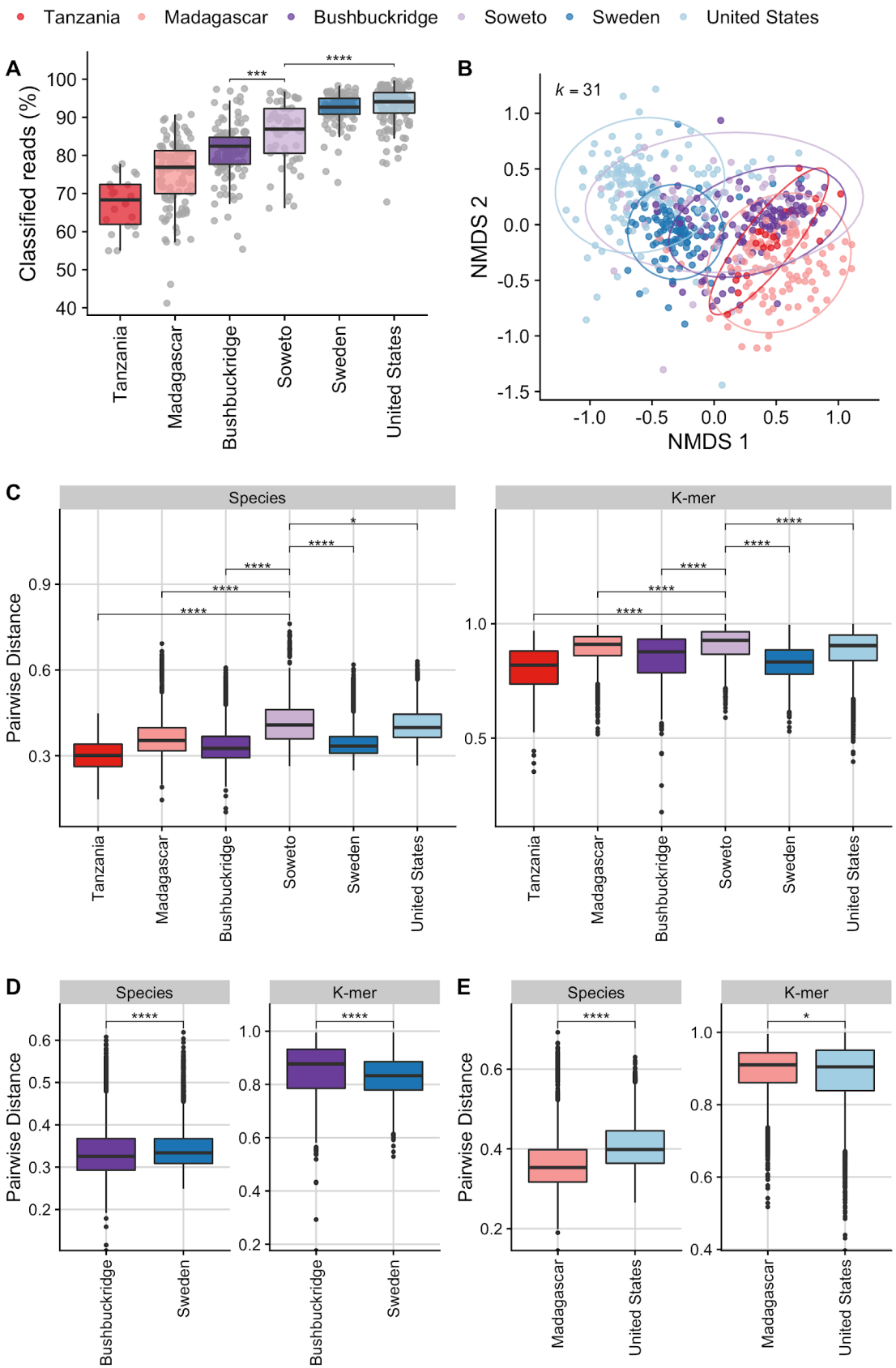
Figure 3. Community-level comparison of global microbiomes

Comparisons of South African microbiome data to microbiome sequence data from four publicly available cohorts representing western (United States, Sweden) and nonwestern (Hadza hunter-gatherers of Tanzania, rural Madagascar) populations.

(A) Number of participants per cohort.

(B) Multidimensional scaling of pairwise Bray-Curtis distance between samples from six datasets of healthy adult shotgun microbiome sequencing data. Western populations (Sweden, United States) cluster away from African populations practicing a traditional lifestyle (Madagascar, Tanzania) while transitional South African microbiomes

798 overlap with both western and nonwestern populations. Shown below are scatterplots of relative abundance of the
799 top four taxa most correlated with MDS 1 (Spearman's rho, *Streptomycetaceae* 0.853, *Spirochaetaceae* 0.850,
800 *Succinivibrionaceae* 0.845, *Bacteroidaceae* -0.801) against MDS 1 on the x axis.
801 (C) Boxplot of the first axis of MDS (MDS 1) which correlates with geography and lifestyle, and the second axis of
802 MDS (MDS 2) where South African populations display a shift relative to other cohorts.
803 (D) Shannon diversity across cohorts. Shannon diversity was calculated from data rarefied to the number of
804 sequence reads of the lowest sample.


Tanzania
Madagascar
Bushbuckridge
Soweto
Sweden
United States
A
Classified reads (%)
B
k = 31
NMDS 2
NMDS 1
C
Species
K-mer
Pairwise Distance
D
E

805

806 **Figure 4. Comparison of beta diversity between communities calculated by taxonomy versus nucleotide *k*-mer**
807 **composition**
808 (A) Percentage of reads classified at any taxonomic rank, by cohort, based on a reference database of all scaffold or
809 higher quality reference genomes in GenBank and RefSeq as of January 2020. Western microbiomes have a higher
810 percentage of classifiable reads compared to nonwestern microbiomes (Wilcoxon rank-sum test $p < 0.001$).
811 (B) Nucleotide sequences of microbiome sequencing reads were compared using *k*-mer sketches. This reference-
812 free approach is not constrained by comparison to existing genomes and therefore allows direct comparison of
813 sequences. Briefly, a hash function generates signatures at varying sequence lengths (*k*) and *k*-mer sketches can be
814 compared between samples. Data shown here are generated from comparisons at *k*=31 (approx. species-
815 level)(Koslicki and Falush, 2016). Non-metric multidimensional scaling (NMDS) of angular distance values
816 computed between each pair of samples.
817 (C-E) Comparison of pairwise beta diversity within communities assessed by Bray-Curtis distance based on
818 species-level classifications and angular distance of nucleotide *k*-mer sketches. (C) All populations. (D) South
819 African populations (Bushbuckridge and Soweto) compared to the Swedish cohort. Beta diversity measured by
820 Bray-Curtis distance is higher in Soweto but lower in Bushbuckridge compared to the United States. However,
821 reference-independent *k*-mer comparisons indicate that nucleotide dissimilarity is higher within both South African
822 populations compared to the Swedish cohort. (E) Species-based Bray-Curtis distance indicates that there is more
823 beta diversity within the United States cohort compared to Malagasy, but *k*-mer distance indicates an opposite
824 pattern.
825 Significance values for Wilcoxon rank sum tests denoted as follows: (*) $p < 0.05$, (**) $p < 0.01$, (***) $p < 0.001$,
826 (****) $p < 0.0001$.

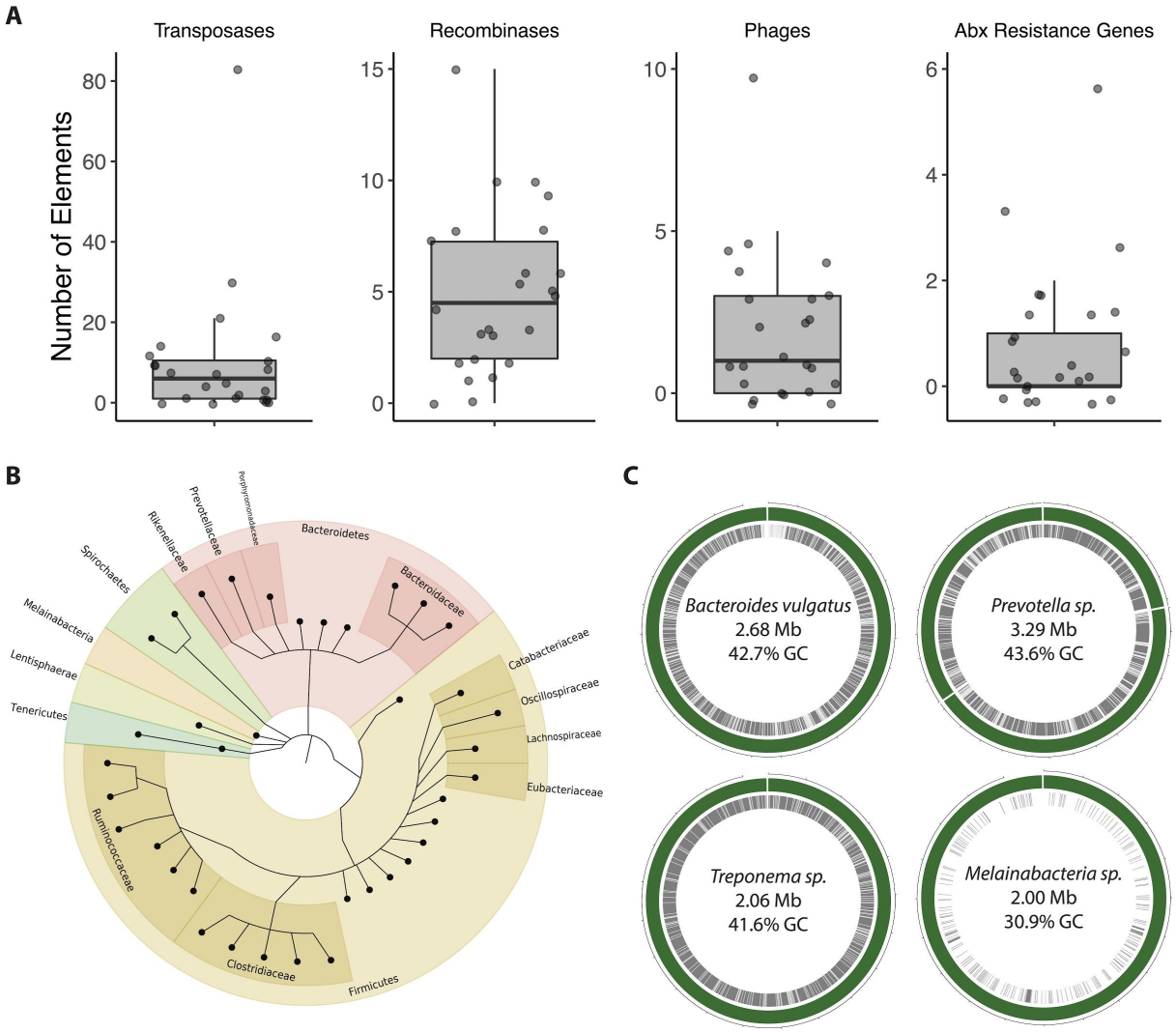


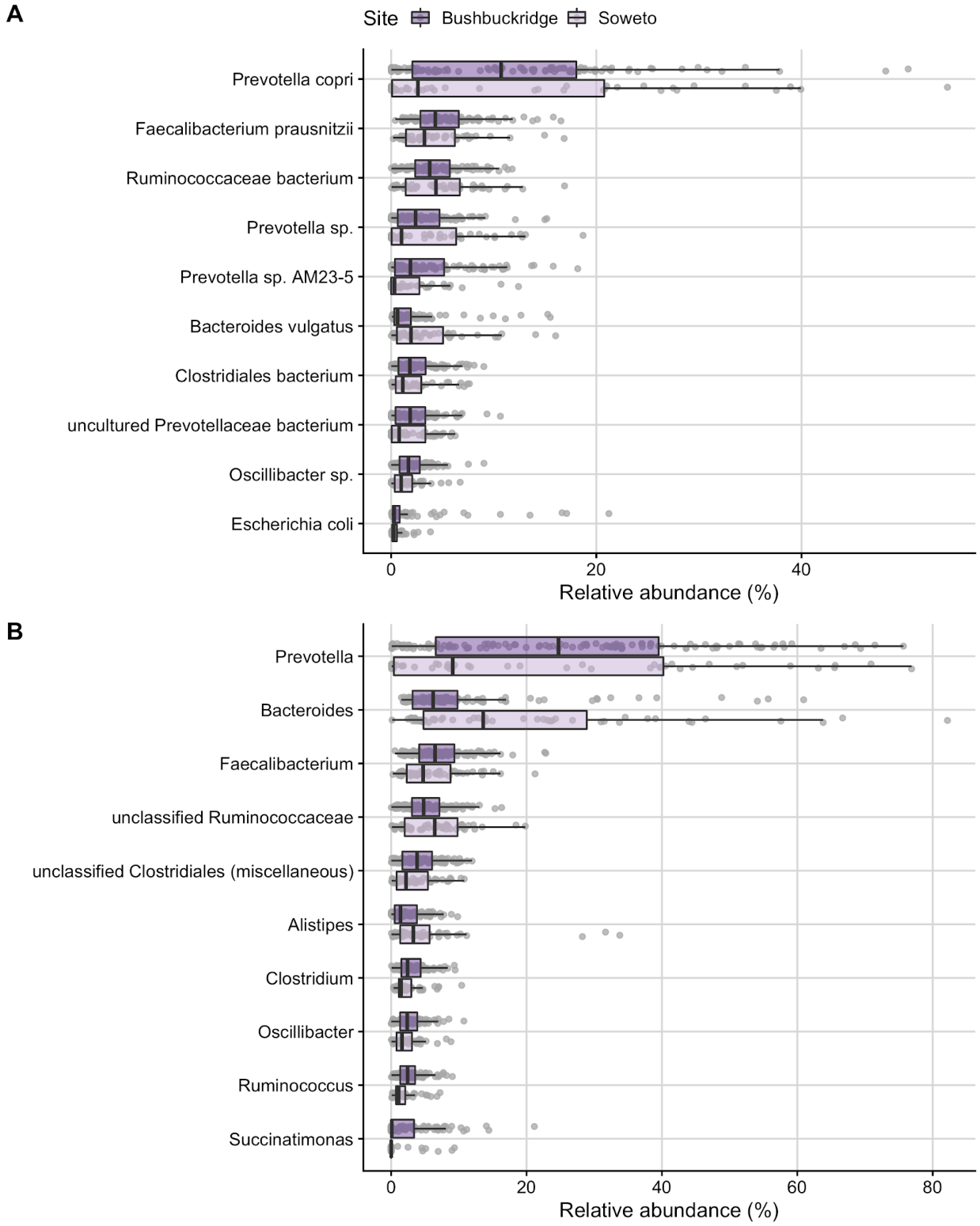
Figure 5. Complete and contiguous genomes of South African microbiota

(A) Number of genomic elements present in medium- and high-quality nanopore MAGs that are absent in corresponding short-read MAGs for the same organism.

(B) Taxonomic classification of de-replicated medium- and high-quality nanopore MAGs. Larger circles represent nanopore MAGs, at the highest level of taxonomic classification.

(C) A selection of MAGs assembled from long-read sequencing (green) of three South African samples compared contigs assembled from corresponding short read data (grey). Outer light grey ring indicates contig scale, with ticks at 100kb intervals. Breaks in circles represent different contigs.

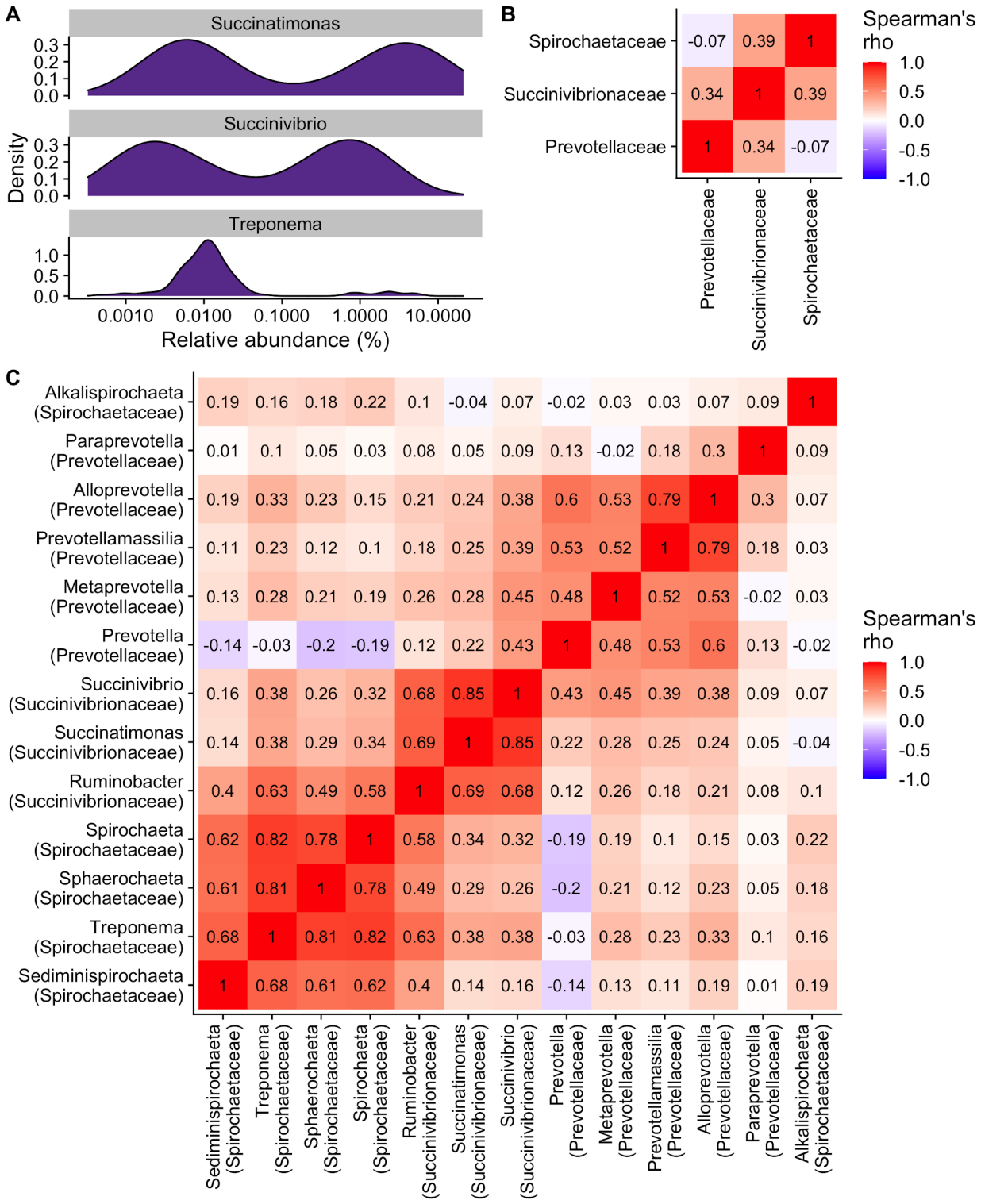
836 # Supplementary Figures



837

838 **Supplementary Figure 1. Most abundant species and genera**

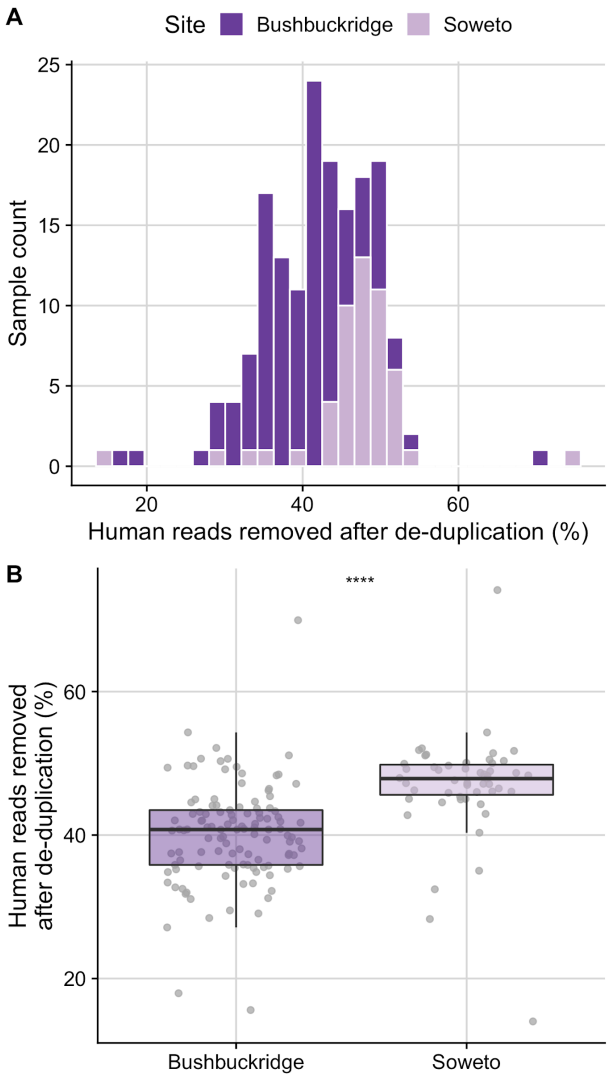
839 Most abundant taxa by mean relative abundance (total sum scaling) shown for samples from
840 Bushbuckridge (n=117) and Soweto (n=51). Taxa are plotted in decreasing order of mean
841 relative abundance calculated across both cohorts combined. Upper and lower box plot whiskers
842 represent the highest and lowest values within 1.5 times the interquartile range, respectively.
843 (A) The most abundant species are *Prevotella copri*, *Faecalibacterium prausnitzii,* and a
844 bacterium from the family Ruminococcaceae.
845 (B) *Prevotella*, *Bacteroides,* and *Faecalibacterium* are the most abundant genera across both
846 study sites.



Supplementary Figure 2. Bimodal distribution of three VANISH taxa

(A) *Succinatimonas, Succinivibrio*, and *Treponema* relative abundance values follow a bimodal distribution in Bushbuckridge.

851 Across all South African samples, several VANISH families (B) and genera (C) are correlated,
852 with the exception of *Prevotella* and genera of the family *Spirochaetaceae* which are not
853 correlated with *Prevotella* (*Treponema*) or weakly anti-correlated with *Prevotella* (*Spirochaeta*,
854 *Sphaerochaeta*, *Sediminispirochaeta*).



855

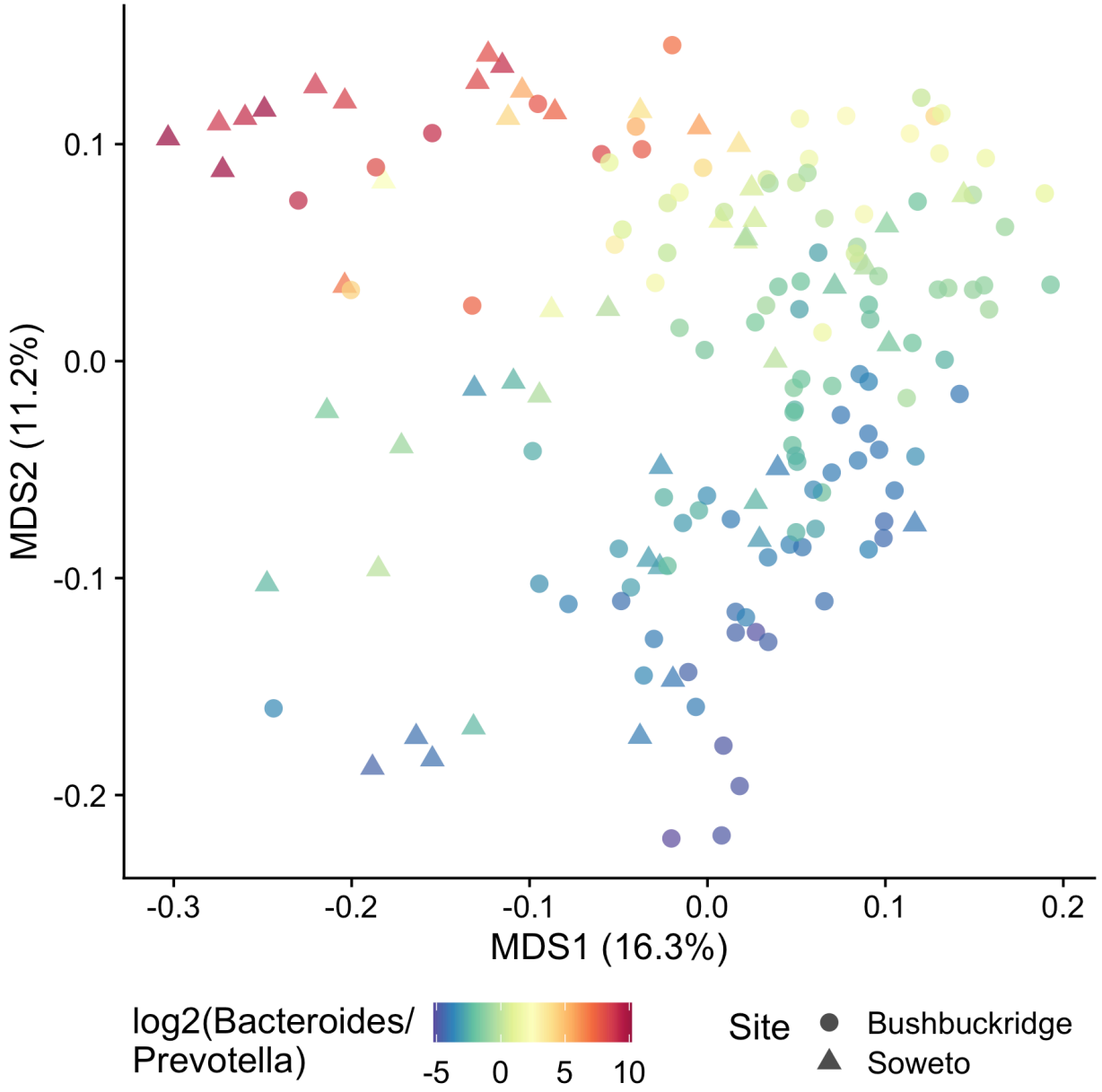
856 **Supplementary Figure 3. Abundance of human reads in metagenomic sequencing**

857 (A) Histogram and (B) box and whisker plots indicating that the proportion of human reads

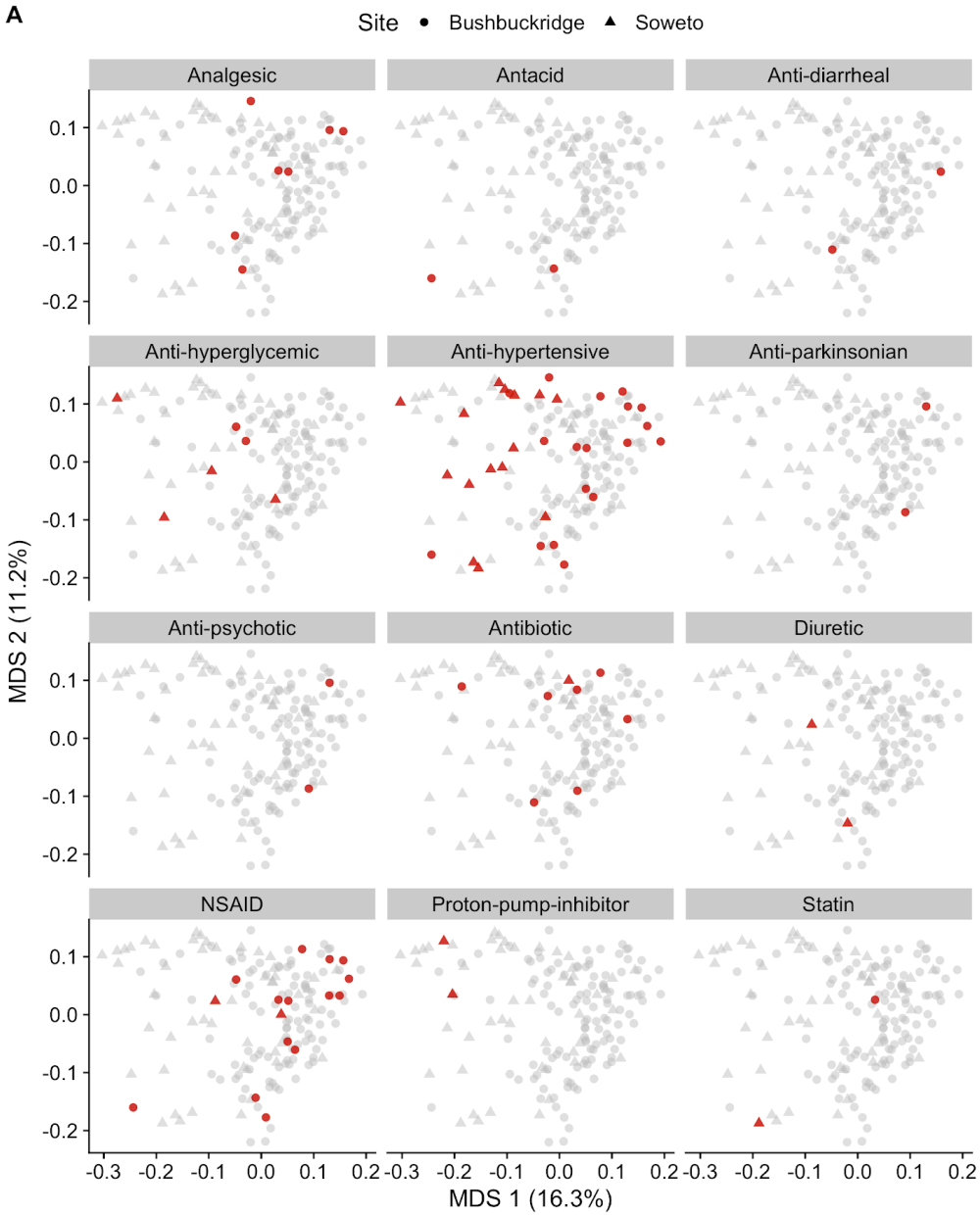
858 removed after deduplication was found to be higher in the Soweto cohort compared to

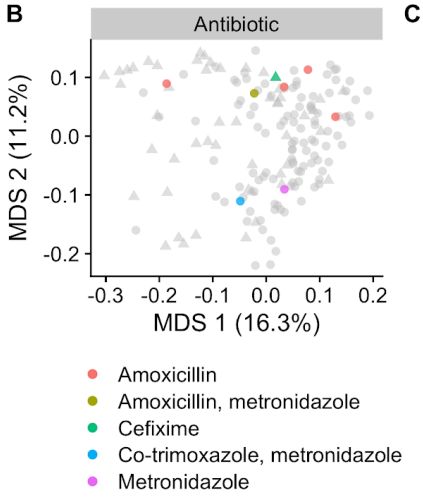
859 Bushbuckridge.

860



Supplementary Figure 4. Bacteroides/Prevotella gradient across study population
Multidimensional scaling ordination of Bray-Curtis distance calculated from species classifications in South African microbiome samples colored by log2 ratio of the relative abundance of the genera *Bacteroides Prevotella*. *Bacteroides* and *Prevotella* are major axes of variation across study samples.





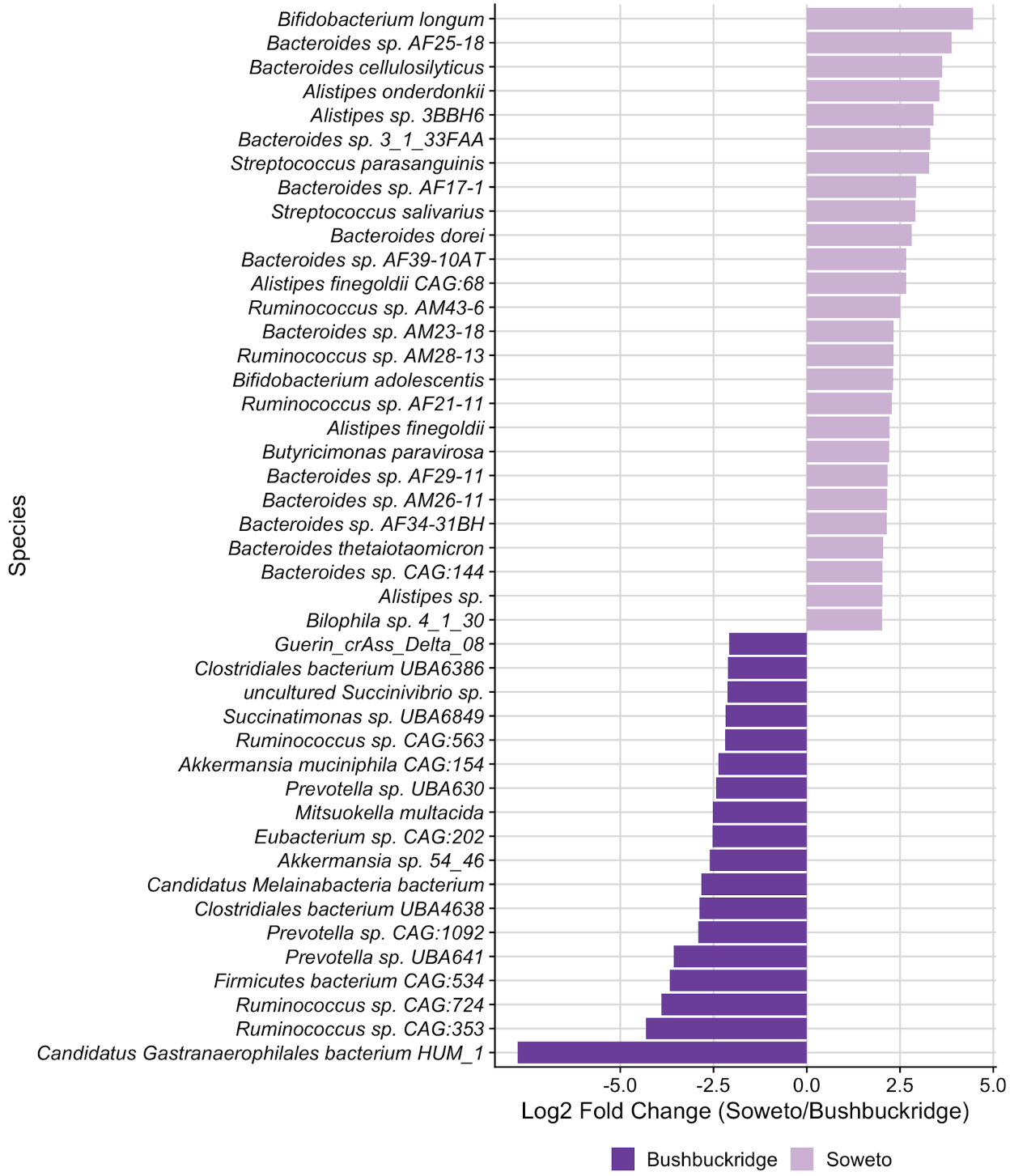
C

Category	R2	Pr(>F)
Analgesic	0.0054	0.504
Antacid	0.0053	0.459
Anti-diarrheal	0.0050	0.555
Anti-hyperglycemic	0.0096	0.070
Anti-hypertensive	0.0048	0.738
Anti-parkinsonian	0.0044	0.725
Anti-psychotic	0.0044	0.736
Antibiotic	0.0045	0.814
Diuretic	0.0059	0.381
NSAID	0.0065	0.318
Proton-pump-inhibitor	0.0136	0.026
Statin	0.0056	0.411

868

869 **Supplementary Figure 5: Concomitant medications do not substantially impact community**
870 **composition**

871 Multidimensional scaling ordination of Bray-Curtis distance calculated from species
872 classifications. Circles indicate participants from Bushbuckridge, triangles indicate participants
873 from Soweto.
874 (A) Points are colored red if the participant was taking a medication of the corresponding class,
875 patients not taking a medication of that class are shown in gray.
876 (B) Specific antibiotics taken by participants. Points are colored according to the antibiotic or
877 combination of antibiotics reported.
878 (C) PERMANOVA R^2 values and p-values for the variation explained by each drug class.



879

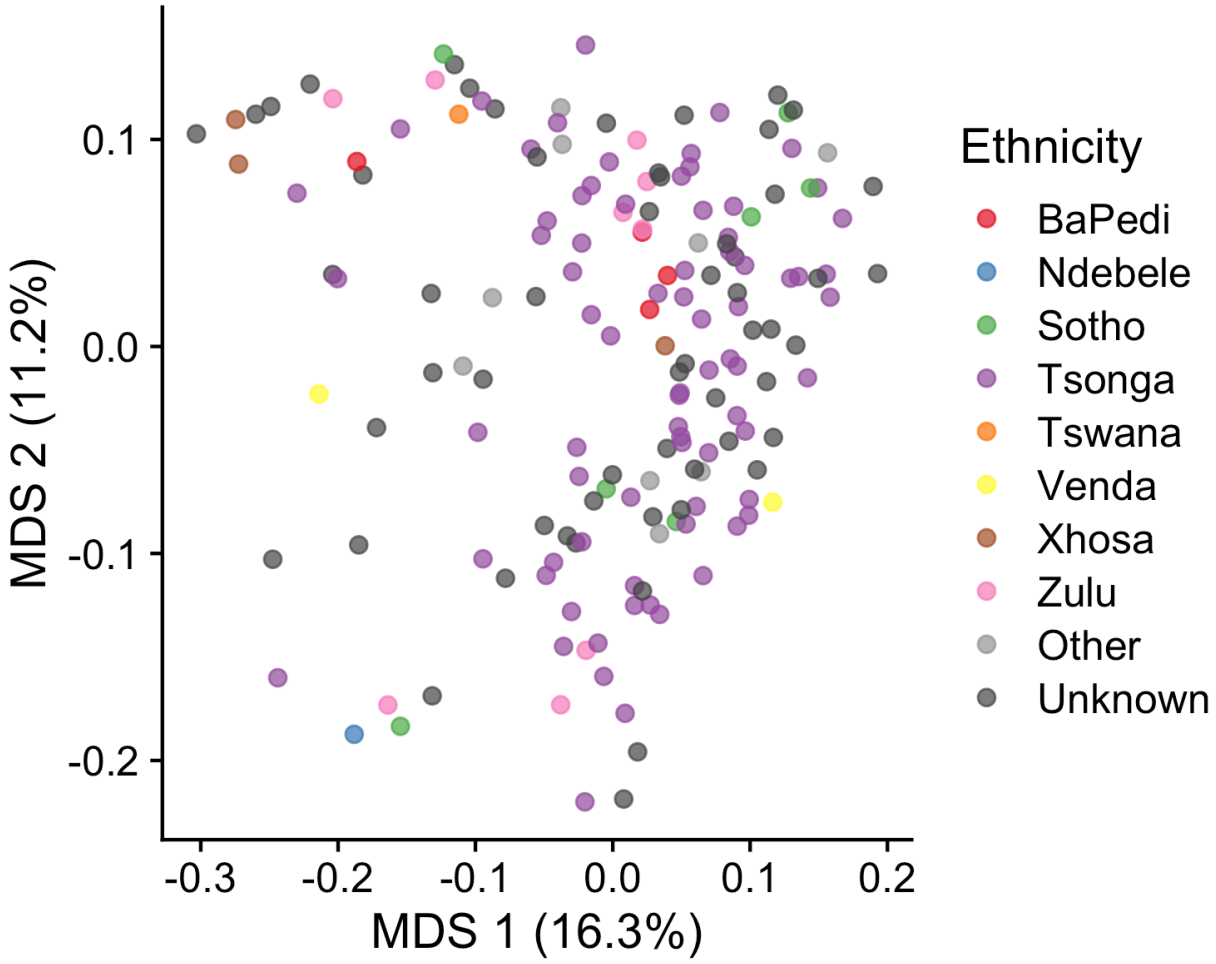
880 **Supplementary Figure 6. Differentially abundant species between Bushbuckridge and**

881 **Soweto**

882 Differentially abundant microbial species between rural Bushbuckridge and urban Soweto

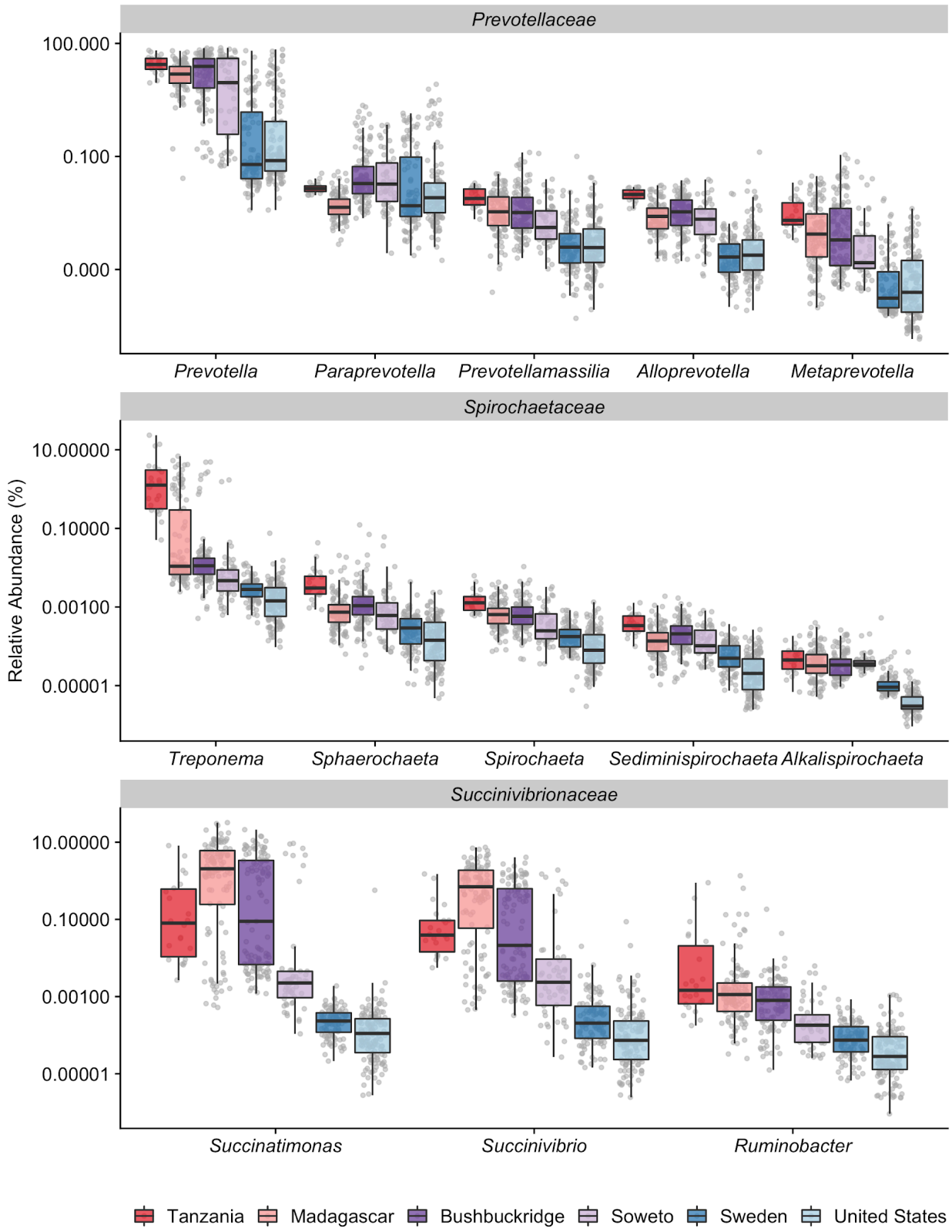
883 samples identified by DESeq2. Features with log2 fold change greater than one are shown (full

884 results in Table S7). Note that differentially abundant microbial genera are presented in Figure
885 2c.



Supplementary Figure 7. South African microbiomes do not cluster by self-reported ethnicity

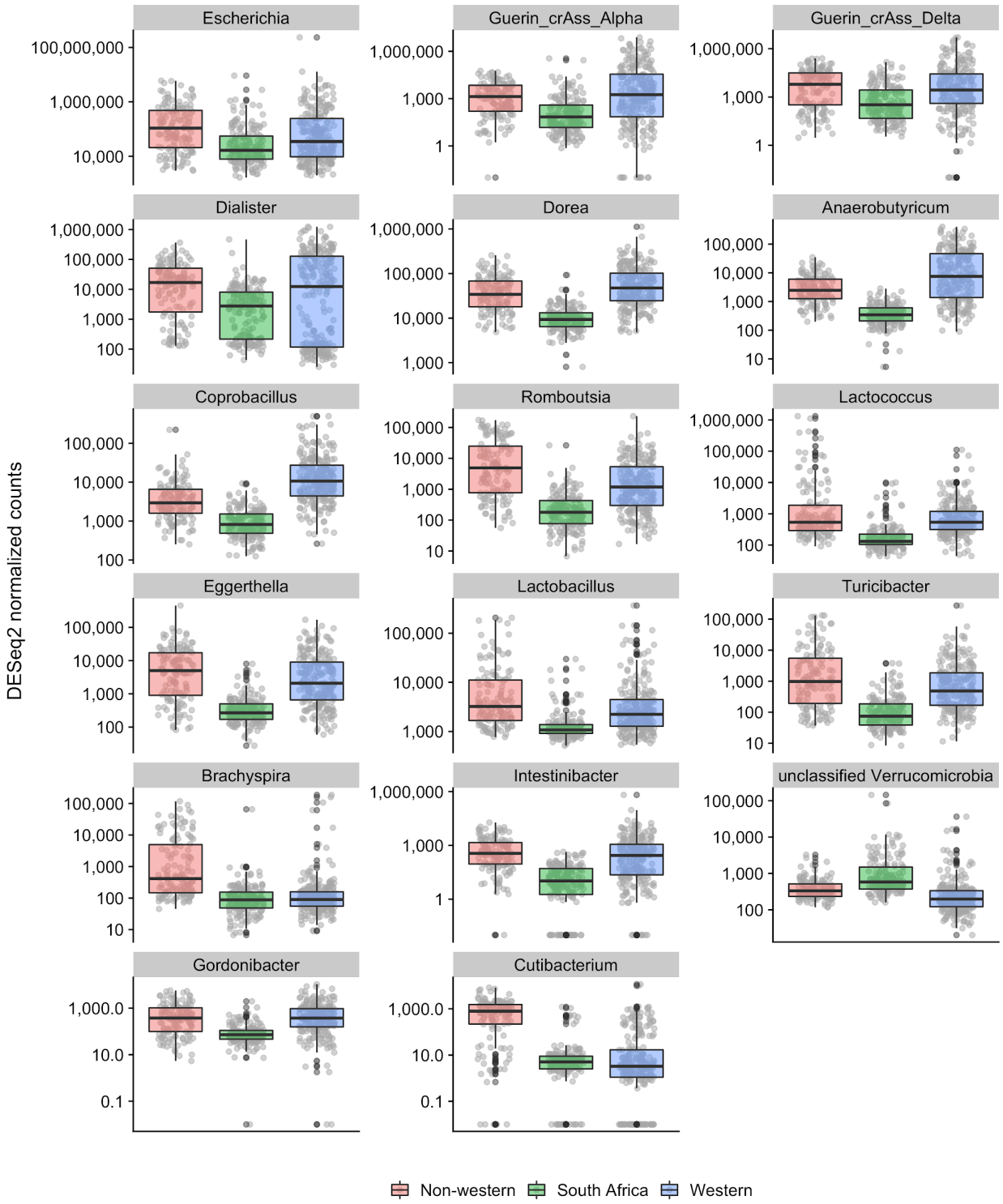
Multidimensional scaling ordination of Bray-Curtis distance with samples are colored by self-reported ethnicity. Samples do not cluster by self-reported ethnicity.



Supplementary Figure 8. Relative abundance of VANISH taxa in global cohort

Relative abundance of VANISH genera from the families Prevotellaceae, Spirochaetaceae, and Succinivibrionaceae. A pseudocount of 1 read was added to each sample prior to relative

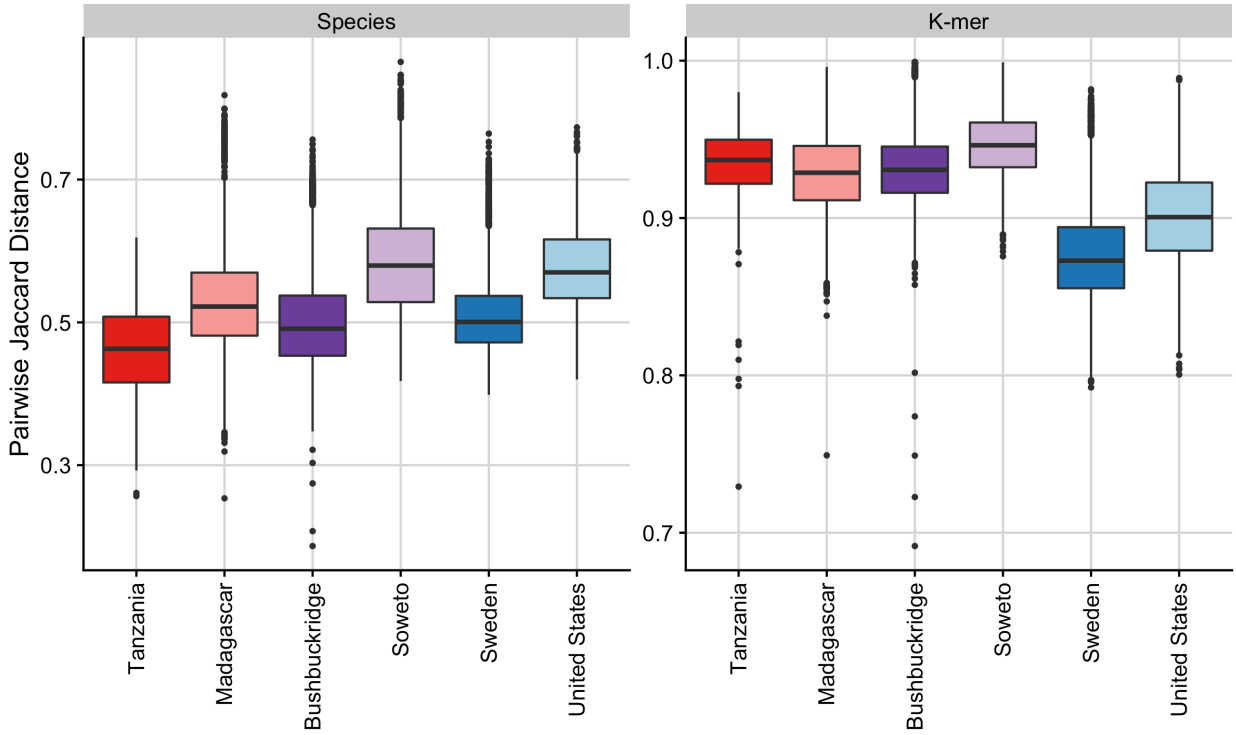
abundance normalization in order to plot on a log scale. Relative abundance values for most genera trend toward decreasing from nonwestern cohorts to western cohorts.



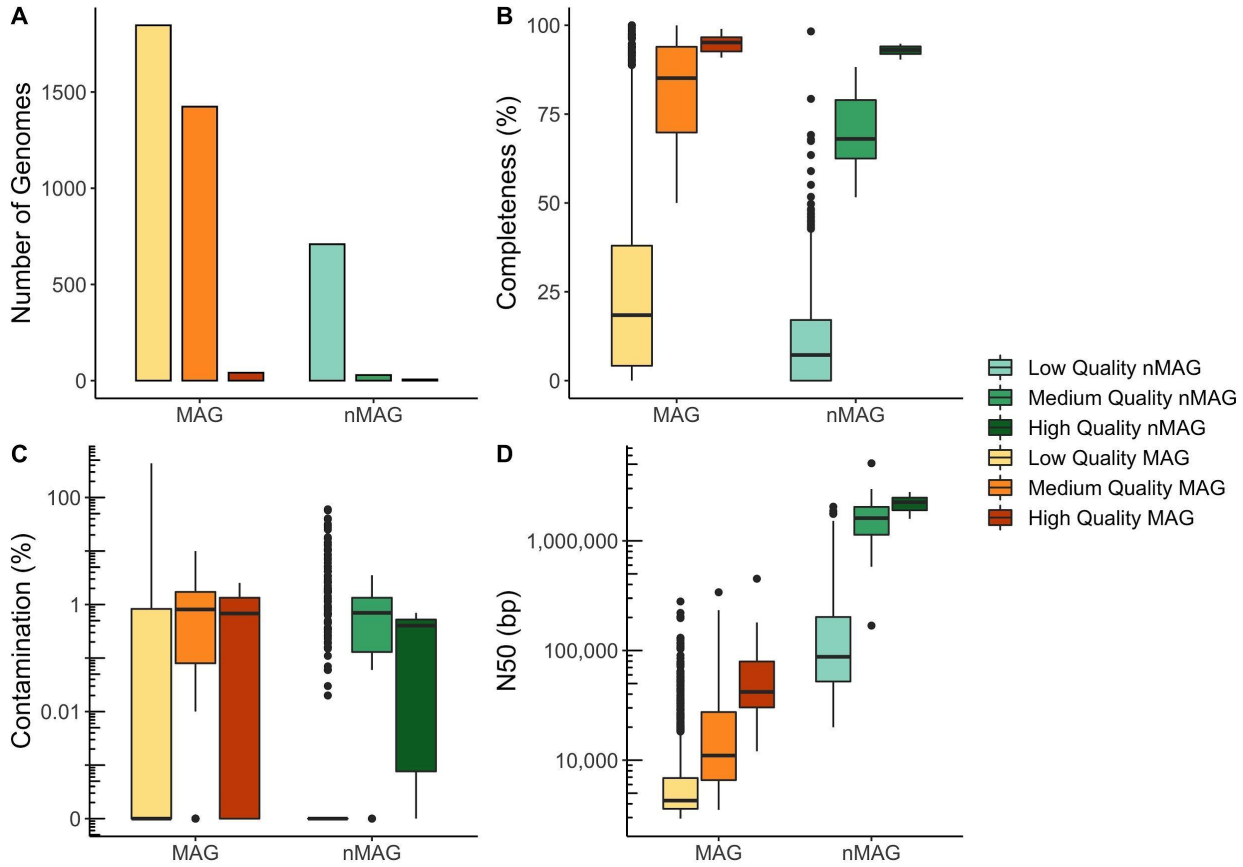
Supplementary Figure 9. Microbial genera which distinguish Bushbuckridge and Soweto

Samples were grouped by geographic region into "western" (USA, Sweden), "nonwestern" (Tanzania, Madagascar) and "South African" (Bushbuckridge, Soweto) and taxa which distinguish the South African group from the western and nonwestern groups were determined separately using DESeq2. Results with the same directionality of log2 fold change with respect

903 to South Africa in both comparisons, with a minimum log2 fold change of 2 in each comparison,
904 are shown. A pseudo-count was added to zero values for plotting.



Supplementary Figure 10. Cohort-wise beta diversity computed via Jaccard distance Comparison of pairwise beta diversity within each cohort based on Jaccard distance between species abundance counts and nucleotide *k*-mer sketches. Nonwestern populations have greater beta diversity than western populations considering nucleotide *k*-mer composition.



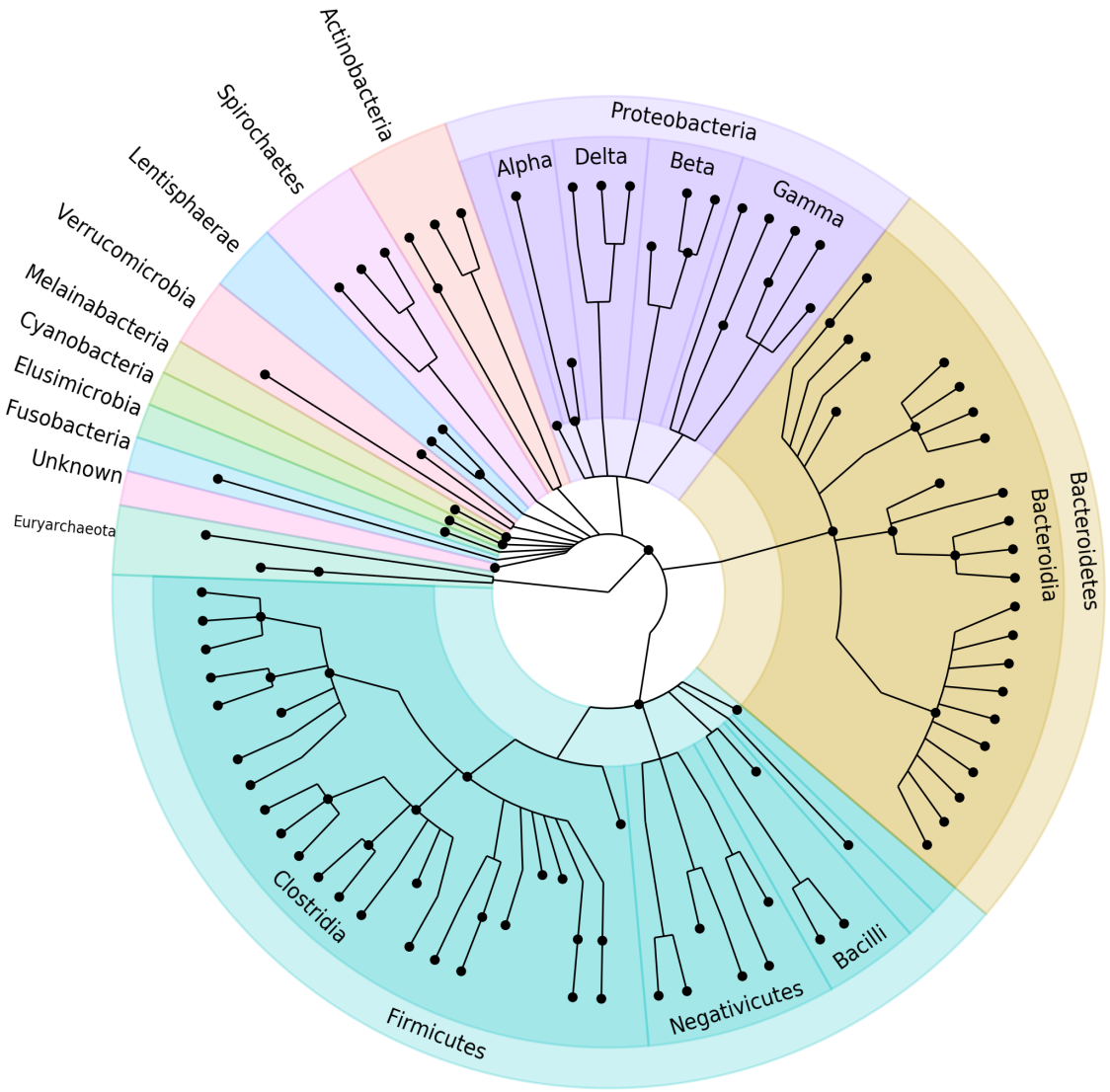
Supplementary Figure 11. Summary statistics for Illumina and nanopore MAGs generated from all samples.

(A) Number of low-, medium-, and high-quality genomes as evaluated with Bowers et al. standards

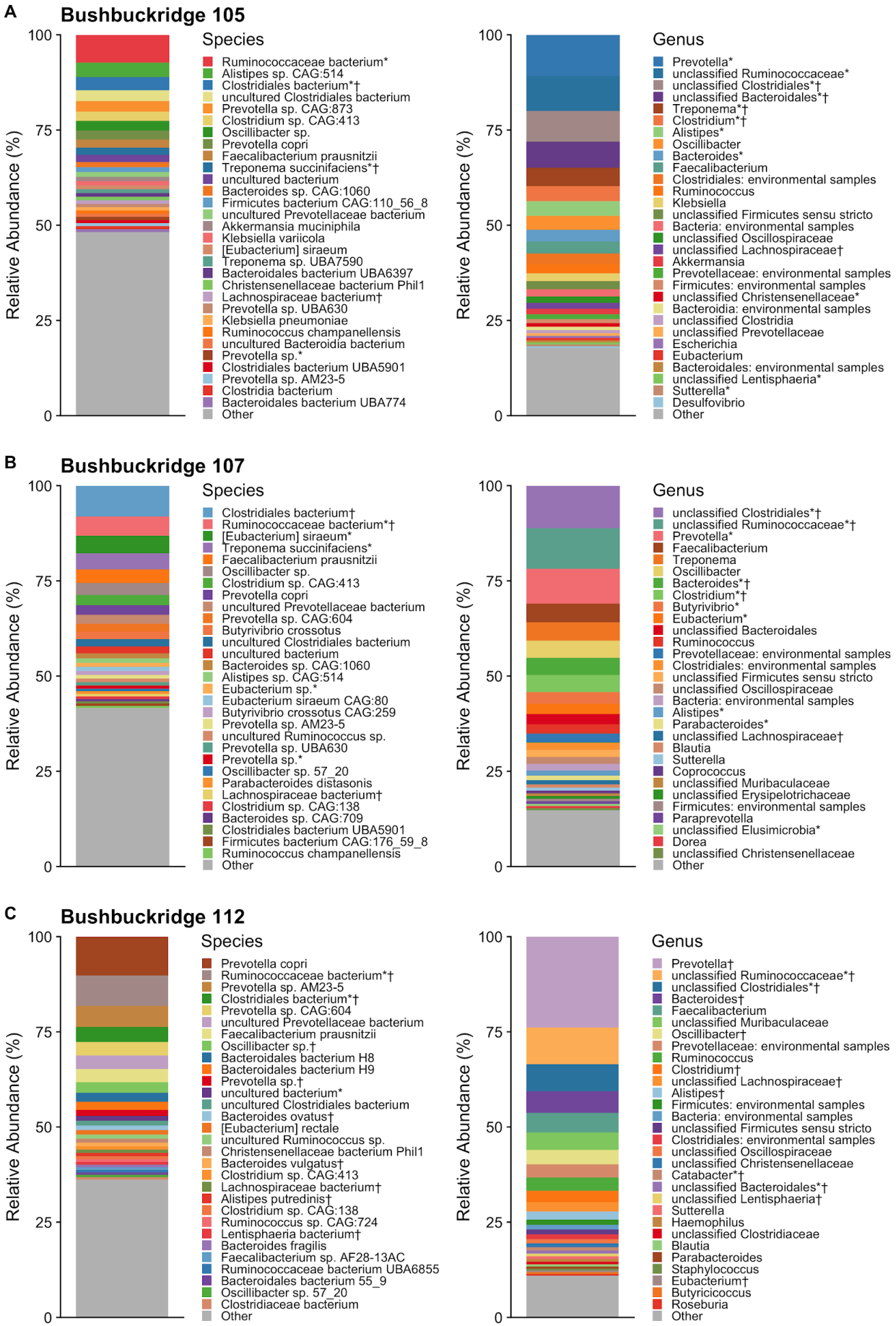
(B) Distribution of MAG percent completeness as determined by CheckM.

(C) Distribution of MAG percent contamination as determined by CheckM.

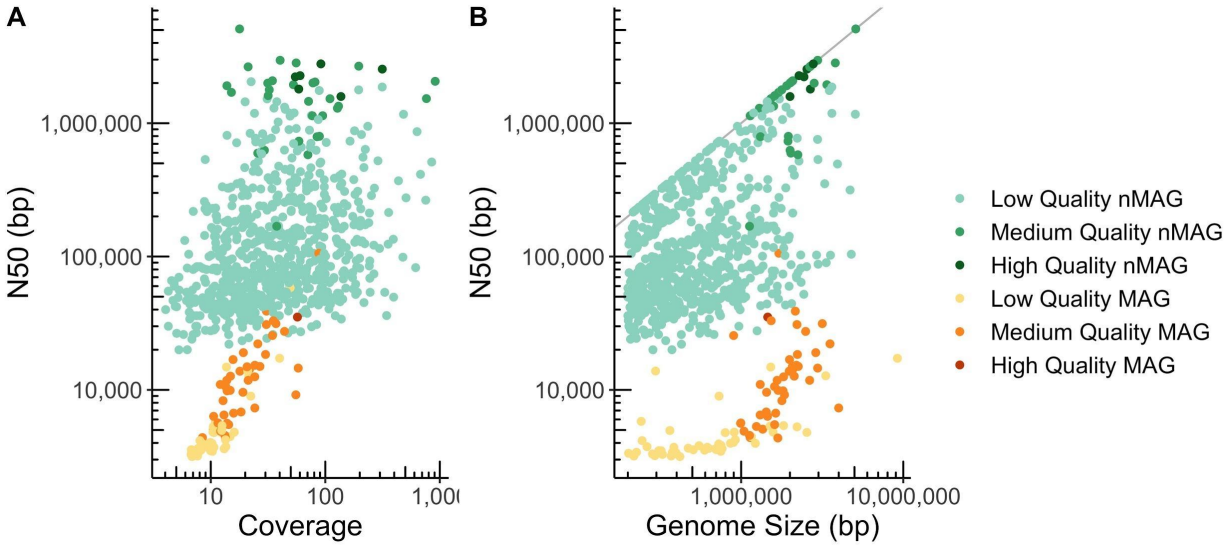
(D) Distribution of MAG N50.



Supplementary Figure 12. Taxonomy of de-replicated Illumina MAGs from all samples
Taxonomic classification of de-replicated medium- and high-quality Illumina MAGs, where black dots indicate a MAG assembled at that level of taxonomic classification. Multiple MAGs at the same classification level are collapsed into single points.


A
Bushbuckridge 105
Relative Abundance (%)
100
75
50
25
0
Species
Ruminococcaceae bacterium*
Alistipes sp. CAG:514
Clostridiales bacterium*†
uncultured Clostridiales bacterium
Prevotella sp. CAG:873
Clostridium sp. CAG:413
Oscillibacter sp.
Prevotella copri
Faecalibacterium prausnitzii
Treponema succinifaciens*†
uncultured bacterium
Bacteroides sp. CAG:1060
Firmicutes bacterium CAG:110_56_8
uncultured Prevotellaceae bacterium
Akkermansia muciniphila
Klebsiella variicola
[Eubacterium] siraeum
Treponema sp. UBA7590
Bacteroidales bacterium UBA6397
Christensenellaceae bacterium Phil1
Lachnospiraceae bacterium†
Prevotella sp. UBA630
Klebsiella pneumoniae
Ruminococcus champanellensis
uncultured Bacteroidia bacterium
Prevotella sp.*
Clostridiales bacterium UBA5901
Prevotella sp. AM23-5
Clostridia bacterium
Bacteroidales bacterium UBA774
Other
Genus
Prevotella*
unclassified Ruminococcaceae*
unclassified Clostridiales*†
unclassified Bacteroidales*†
Treponema*†
Clostridium*†
Alistipes*
Oscillibacter
Bacteroides*
Faecalibacterium
Clostridiales: environmental samples
Ruminococcus
Klebsiella
unclassified Firmicutes sensu stricto
Bacteria: environmental samples
unclassified Oscillospiraceae
unclassified Lachnospiraceae†
Akkermansia
Prevotellaceae: environmental samples
Firmicutes: environmental samples
unclassified Christensenellaceae*
Bacteroidia: environmental samples
unclassified Clostridia
unclassified Prevotellaceae
Escherichia
Eubacterium
Bacteroidales: environmental samples
unclassified Lentisphaeria*
Sutterella*
Desulfovibrio
Other
B
Bushbuckridge 107
Relative Abundance (%)
100
75
50
25
0
Species
Clostridiales bacterium†
Ruminococcaceae bacterium*†
[Eubacterium] siraeum*
Treponema succinifaciens*
Faecalibacterium prausnitzii
Oscillibacter sp.
Clostridium sp. CAG:413
Prevotella copri
uncultured Prevotellaceae bacterium
Prevotella sp. CAG:604
Butyrivibrio crossotus
uncultured Clostridiales bacterium
uncultured bacterium
Bacteroides sp. CAG:1060
Alistipes sp. CAG:514
Eubacterium sp.*
Eubacterium siraeum CAG:80
Butyrivibrio crossotus CAG:259
Prevotella sp. AM23-5
uncultured Ruminococcus sp.
Prevotella sp. UBA630
Prevotella sp.*
Oscillibacter sp. 57_20
Parabacteroides distasonis
Lachnospiraceae bacterium†
Clostridium sp. CAG:138
Bacteroides sp. CAG:709
Clostridiales bacterium UBA5901
Firmicutes bacterium CAG:176_59_8
Ruminococcus champanellensis
Other
Genus
unclassified Clostridiales*†
unclassified Ruminococcaceae*†
Prevotella*
Faecalibacterium
Treponema
Oscillibacter
Bacteroides*†
Clostridium*†
Butyrivibrio*
Eubacterium*
unclassified Bacteroidales
Ruminococcus
Prevotellaceae: environmental samples
Clostridiales: environmental samples
unclassified Firmicutes sensu stricto
unclassified Oscillospiraceae
Bacteria: environmental samples
Alistipes*
Parabacteroides*
unclassified Lachnospiraceae†
Blautia
Sutterella
Coprococcus
unclassified Muribaculaceae
unclassified Erysipelotrichaceae
Firmicutes: environmental samples
Paraprevotella
unclassified Elusimicrobia*
Dorea
unclassified Christensenellaceae
Other
C
Bushbuckridge 112
Relative Abundance (%)
100
75
50
25
0
Species
Prevotella copri
Ruminococcaceae bacterium*†
Prevotella sp. AM23-5
Clostridiales bacterium*†
Prevotella sp. CAG:604
uncultured Prevotellaceae bacterium
Faecalibacterium prausnitzii
Oscillibacter sp.†
Bacteroidales bacterium H8
Bacteroidales bacterium H9
Prevotella sp.†
uncultured bacterium*
uncultured Clostridiales bacterium
Bacteroides ovatus†
[Eubacterium] rectale
uncultured Ruminococcus sp.
Christensenellaceae bacterium Phil1
Bacteroides vulgatus†
Clostridium sp. CAG:413
Lachnospiraceae bacterium†
Alistipes putredinis†
Clostridium sp. CAG:138
Ruminococcus sp. CAG:724
Lentisphaeria bacterium†
Bacteroides fragilis
Faecalibacterium sp. AF28-13AC
Ruminococcaceae bacterium UBA6855
Bacteroidales bacterium 55_9
Oscillibacter sp. 57_20
Clostridiaceae bacterium
Other
Genus
Prevotella†
unclassified Ruminococcaceae*†
unclassified Clostridiales*†
Bacteroides†
Faecalibacterium
unclassified Muribaculaceae
Oscillibacter†
Prevotellaceae: environmental samples
Ruminococcus
Clostridium†
unclassified Lachnospiraceae†
Alistipes†
Firmicutes: environmental samples
Bacteria: environmental samples
unclassified Firmicutes sensu stricto
Clostridiales: environmental samples
unclassified Oscillospiraceae
unclassified Christensenellaceae
Catabacter*†
unclassified Bacteroidales*†
unclassified Lentisphaeria†
Sutterella
Haemophilus
unclassified Clostridiaceae
Blautia
Parabacteroides
Staphylococcus
Eubacterium†
Butyricicoccus
Roseburia
Other

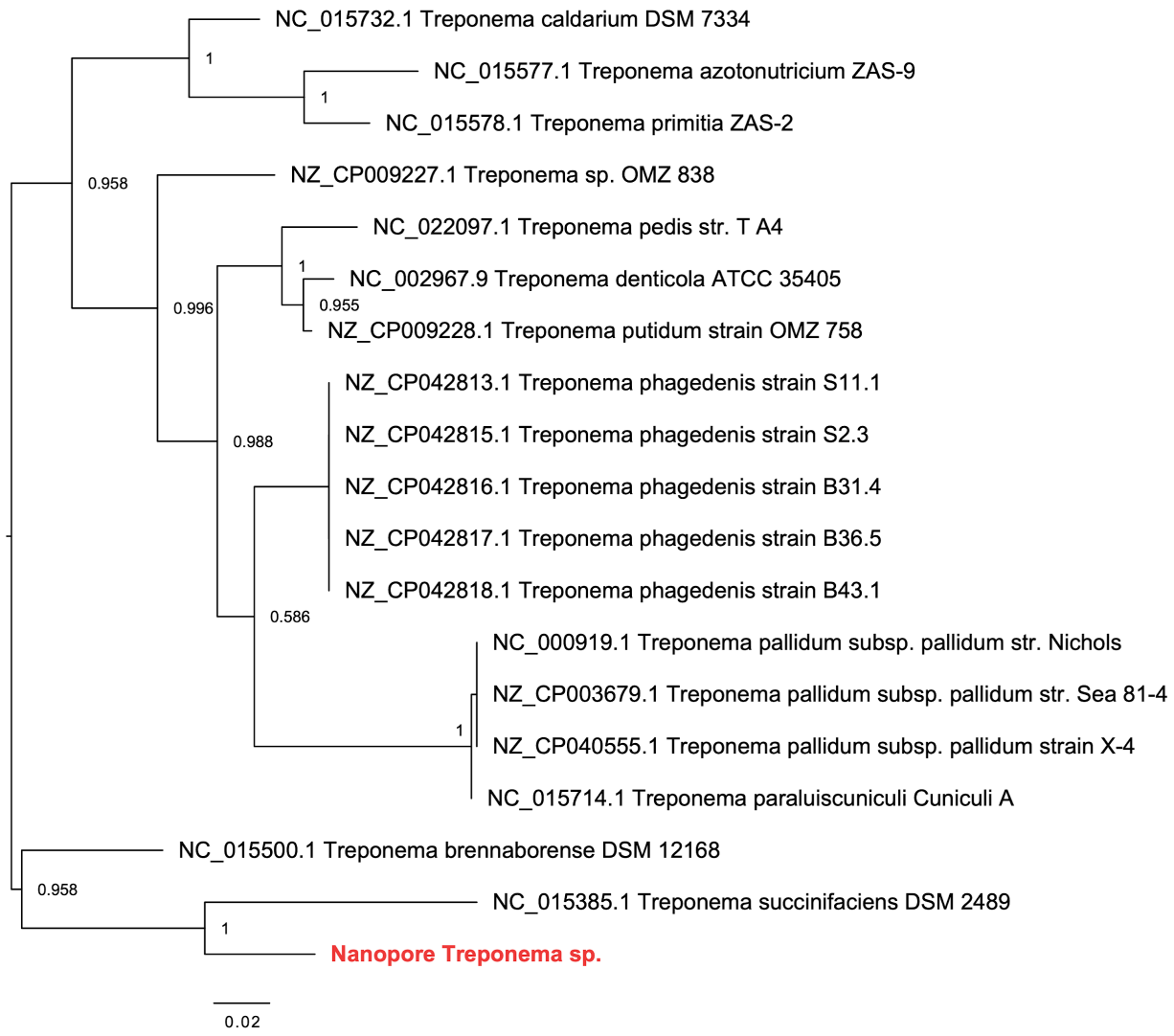
924 **Supplementary Figure 13. Taxonomic composition for samples selected for nanopore**
925 **sequencing**
926 Short-read sequencing-based taxonomic classifications for the three samples selected for
927 Nanopore sequencing, showing (A) genus-level and (B) species-level classifications. Top thirty
928 taxa by relative abundance shown in each plot. Symbols indicate whether a medium- or high-
929 quality short-read (*) or nanopore MAG (†) was assembled from the corresponding genus or
930 species



Supplementary Figure 14. Summary statistics of nanopore and short read MAGs generated for three Bushbuckridge samples

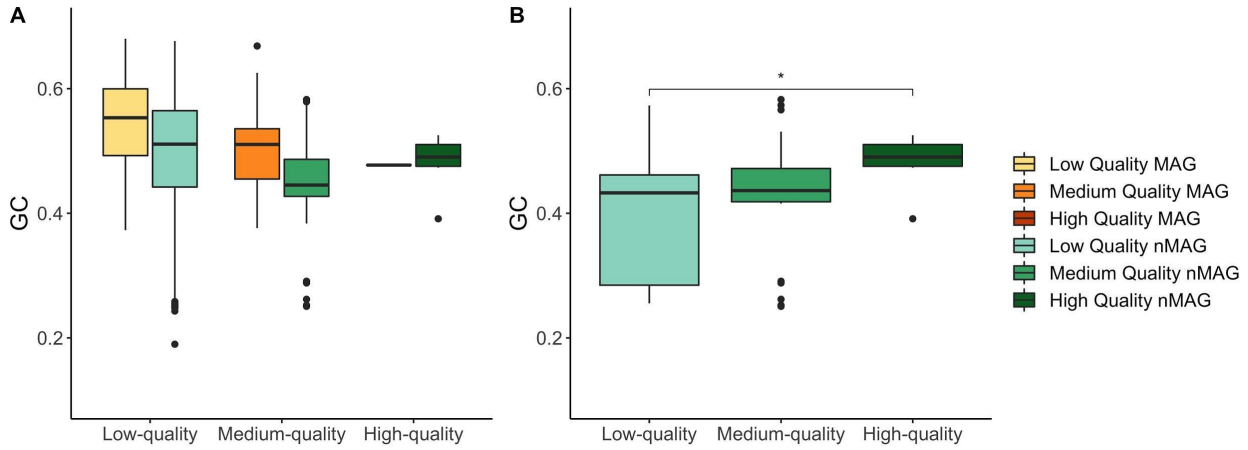
(A) MAG short read or long-read coverage versus MAG N50.

(B) MAG total size versus MAG N50. Grey line indicates where genome N50 equals total genome size.



Supplementary Figure 15. Phylogeny of *Treponema* 16S rRNA sequences

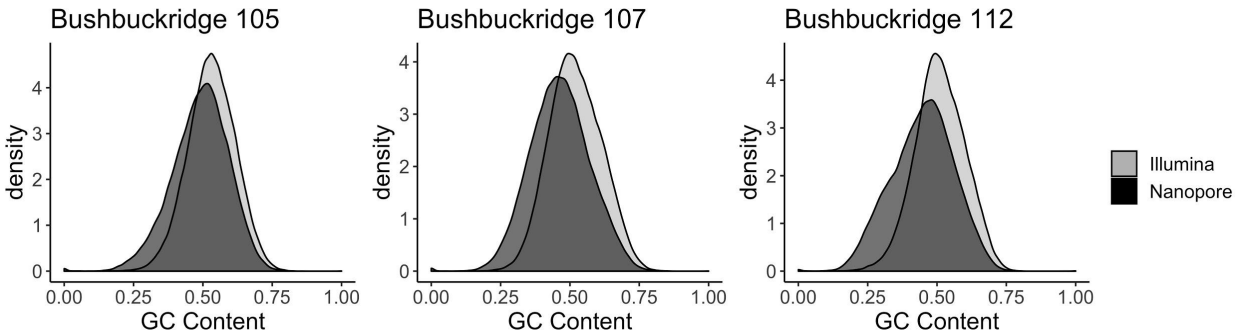
Phylogeny of 16S rRNA sequences from species of the genus *Treponema* show that the *Treponema* sp. assembled via Nanopore sequencing is most related to *T. succinifaciens*, but is phylogenetically distinct. The nanopore genome is highlighted in red font. Branch labels indicate Shimodaira-Hasegawa support values for splits.



Supplementary Figure 16. GC content of MAGs and nMAGs generated from three Bushbuckridge samples

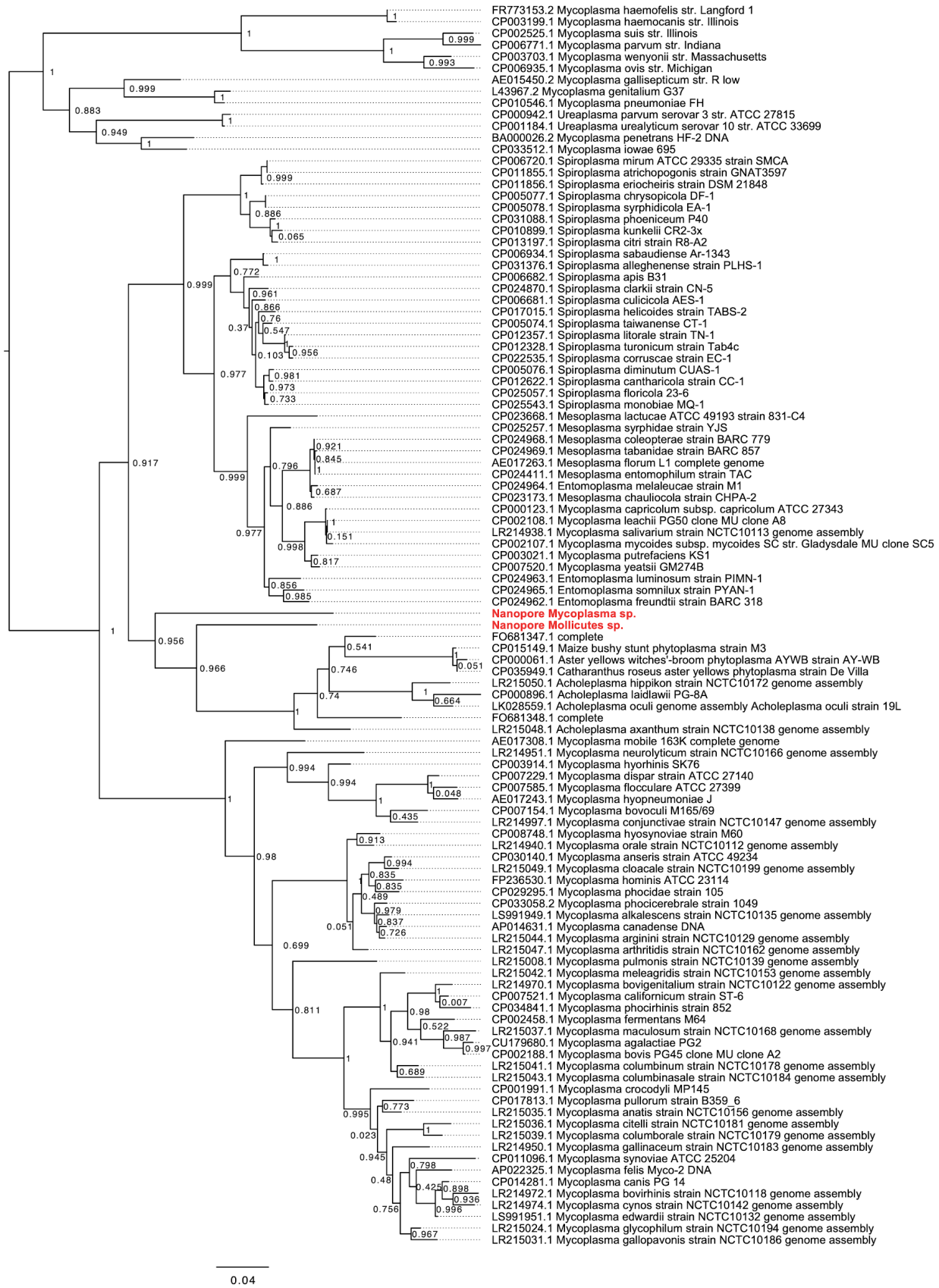
(A) GC content range of MAGs and nMAGs.

(B) nMAGs with contig N50 values greater than one megabase. GC content of low-quality nMAGs is lower than the GC content of high-quality nMAGs, despite nMAGs of all quality having N50 values of higher than one megabase. * = $p \leq 0.05$, Wilcoxon rank sum test.



Supplementary Figure 17. GC content of nanopore and Illumina sequencing reads generated from three Bushbuckridge samples

GC content was calculated for all processed Illumina reads (average length of 126 bp) and for 126 bp windows of all nanopore reads. GC content distribution was subsampled to 100,000 measurements per method.


FR773153.2 Mycoplasma haemofelis str. Langford 1
CP003199.1 Mycoplasma haemocanis str. Illinois
CP002525.1 Mycoplasma suis str. Illinois
CP006771.1 Mycoplasma parvum str. Indiana
CP003703.1 Mycoplasma wenyonii str. Massachusetts
CP006935.1 Mycoplasma ovis str. Michigan
AE015450.2 Mycoplasma gallisepticum str. R low
L43967.2 Mycoplasma genitalium G37
CP010546.1 Mycoplasma pneumoniae FH
CP000942.1 Ureaplasma parvum serovar 3 str. ATCC 27815
CP001184.1 Ureaplasma urealyticum serovar 10 str. ATCC 33699
BA000026.2 Mycoplasma penetrans HF-2 DNA
CP033512.1 Mycoplasma iowae 695
CP006720.1 Spiroplasma mirum ATCC 29335 strain SMCA
CP011855.1 Spiroplasma atrichopogonis strain GNAT3597
CP011856.1 Spiroplasma eriocheiris strain DSM 21848
CP005077.1 Spiroplasma chrysopicola DF-1
CP005078.1 Spiroplasma syrphidicola EA-1
CP031088.1 Spiroplasma phoeniceum P40
CP010899.1 Spiroplasma kunkelii CR2-3x
CP013197.1 Spiroplasma citri strain R8-A2
CP006934.1 Spiroplasma sabaudiense Ar-1343
CP031376.1 Spiroplasma alleghenense strain PLHS-1
CP006682.1 Spiroplasma apis B31
CP024870.1 Spiroplasma clarkii strain CN-5
CP006681.1 Spiroplasma culicicola AES-1
CP017015.1 Spiroplasma helicoides strain TABS-2
CP005074.1 Spiroplasma taiwanense CT-1
CP012357.1 Spiroplasma litorale strain TN-1
CP012328.1 Spiroplasma turonicum strain Tab4c
CP022535.1 Spiroplasma corruscae strain EC-1
CP005076.1 Spiroplasma diminutum CUAS-1
CP012622.1 Spiroplasma cantharicola strain CC-1
CP025057.1 Spiroplasma floricola 23-6
CP025543.1 Spiroplasma monobiae MQ-1
CP023668.1 Mesoplasma lactucae ATCC 49193 strain 831-C4
CP025257.1 Mesoplasma syrphidae strain YJS
CP024968.1 Mesoplasma coleopterae strain BARC 779
CP024969.1 Mesoplasma tabanidae strain BARC 857
AE017263.1 Mesoplasma florum L1 complete genome
CP024411.1 Mesoplasma entomophilum strain TAC
CP024964.1 Entomoplasma melaleucae strain M1
CP023173.1 Mesoplasma chauliocola strain CHPA-2
CP000123.1 Mycoplasma capricolum subsp. capricolum ATCC 27343
CP002108.1 Mycoplasma leachii PG50 clone MU clone A8
LR214938.1 Mycoplasma salivarium strain NCTC10113 genome assembly
CP002107.1 Mycoplasma mycoides subsp. mycoides SC str. Gladysdale MU clone SC5
CP003021.1 Mycoplasma putrefaciens KS1
CP007520.1 Mycoplasma yeatsii GM274B
CP024963.1 Entomoplasma luminosum strain PIMN-1
CP024965.1 Entomoplasma somnilux strain PYAN-1
CP024962.1 Entomoplasma freundtii strain BARC 318
Nanopore Mycoplasma sp.
Nanopore Mollicutes sp.
FO681347.1 complete
CP015149.1 Maize bushy stunt phytoplasma strain M3
CP000061.1 Aster yellows witches'-broom phytoplasma AYWB strain AY-WB
CP035949.1 Catharanthus roseus aster yellows phytoplasma strain De Villa
LR215050.1 Acholeplasma hippikon strain NCTC10172 genome assembly
CP000896.1 Acholeplasma laidlawii PG-8A
LK028559.1 Acholeplasma oculi genome assembly Acholeplasma oculi strain 19L
FO681348.1 complete
LR215048.1 Acholeplasma axanthum strain NCTC10138 genome assembly
AE017308.1 Mycoplasma mobile 163K complete genome
LR214951.1 Mycoplasma neurolyticum strain NCTC10166 genome assembly
CP003914.1 Mycoplasma hyorhinis SK76
CP007229.1 Mycoplasma dispar strain ATCC 27140
CP007585.1 Mycoplasma flocculare ATCC 27399
AE017243.1 Mycoplasma hyopneumoniae J
CP007154.1 Mycoplasma bovoculi M165/69
LR214997.1 Mycoplasma conjunctivae strain NCTC10147 genome assembly
CP008748.1 Mycoplasma hyosynoviae strain M60
LR214940.1 Mycoplasma orale strain NCTC10112 genome assembly
CP030140.1 Mycoplasma anseris strain ATCC 49234
LR215049.1 Mycoplasma cloacale strain NCTC10199 genome assembly
FP236530.1 Mycoplasma hominis ATCC 23114
CP029295.1 Mycoplasma phocidae strain 105
CP033058.2 Mycoplasma phocicerebrale strain 1049
LS991949.1 Mycoplasma alkalescens strain NCTC10135 genome assembly
AP014631.1 Mycoplasma canadense DNA
LR215044.1 Mycoplasma arginini strain NCTC10129 genome assembly
LR215047.1 Mycoplasma arthritidis strain NCTC10162 genome assembly
LR215008.1 Mycoplasma pulmonis strain NCTC10139 genome assembly
LR215042.1 Mycoplasma meleagridis strain NCTC10153 genome assembly
LR214970.1 Mycoplasma bovigenitalium strain NCTC10122 genome assembly
CP007521.1 Mycoplasma californicum strain ST-6
CP034841.1 Mycoplasma phocirhinis strain 852
CP002458.1 Mycoplasma fermentans M64
LR215037.1 Mycoplasma maculosum strain NCTC10168 genome assembly
CU179680.1 Mycoplasma agalactiae PG2
CP002188.1 Mycoplasma bovis PG45 clone MU clone A2
LR215041.1 Mycoplasma columbinum strain NCTC10178 genome assembly
LR215043.1 Mycoplasma columbinasale strain NCTC10184 genome assembly
CP001991.1 Mycoplasma crocodyli MP145
CP017813.1 Mycoplasma pullorum strain B359_6
LR215035.1 Mycoplasma anatis strain NCTC10156 genome assembly
LR215036.1 Mycoplasma citelli strain NCTC10181 genome assembly
LR215039.1 Mycoplasma columborale strain NCTC10179 genome assembly
LR214950.1 Mycoplasma gallinaceum strain NCTC10183 genome assembly
CP011096.1 Mycoplasma synoviae ATCC 25204
AP022325.1 Mycoplasma felis Myco-2 DNA
CP014281.1 Mycoplasma canis PG 14
LR214972.1 Mycoplasma bovirhinis strain NCTC10118 genome assembly
LR214974.1 Mycoplasma cynos strain NCTC10142 genome assembly
LS991951.1 Mycoplasma edwardii strain NCTC10132 genome assembly
LR215024.1 Mycoplasma glycophilum strain NCTC10194 genome assembly
LR215031.1 Mycoplasma gallopavonis strain NCTC10186 genome assembly
0.04

958

959 **Supplementary Figure 18. Phylogeny of Mollicutes 16S rRNA sequences**

960 Phylogeny of 16S rRNA sequences from species of the class Mollicutes showing the Mollicutes
961 and Mycoplasma genomes assembled via nanopore sequencing. Nanopore genomes are
962 highlighted in red font. Branch labels indicate Shimodaira-Hasegawa support values for splits.

References

Ahlmann-Eltze, C. (2019). ggsignif: Significance Brackets for “ggplot2.”

Ajayi, I.O., Adebamowo, C., Adami, H.-O., Dalal, S., Diamond, M.B., Bajunirwe, F., Guwatudde, D., Njelekela, M., Nankya-Mutyoba, J., Chiwanga, F.S., et al. (2016). Urban-rural and geographic differences in overweight and obesity in four sub-Saharan African adult populations: a multi-country cross-sectional study. BMC Public Health *16*, 1126.

Almeida, A., Mitchell, A.L., Boland, M., Forster, S.C., Gloor, G.B., Tarkowska, A., Lawley, T.D., and Finn, R.D. (2019). A new genomic blueprint of the human gut microbiota. Nature *568*, 499–504.

Angelakis, E., Bachar, D., Yasir, M., Musso, D., Djossou, F., Gaborit, B., Brah, S., Diallo, A., Ndombe, G.M., Mediannikov, O., et al. (2019). Treponema species enrich the gut microbiota of traditional rural populations but are absent from urban individuals. New Microbes New Infect *27*, 14–21.

Asnicar, F., Weingart, G., Tickle, T.L., Huttenhower, C., and Segata, N. (2015). Compact graphical representation of phylogenetic data and metadata with GraPhlAn. PeerJ *3*, e1029.

Bäckhed, F., Roswall, J., Peng, Y., Feng, Q., Jia, H., Kovatcheva-Datchary, P., Li, Y., Xia, Y., Xie, H., Zhong, H., et al. (2015). Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. Cell Host Microbe *17*, 690–703.

Bishara, A., Moss, E.L., Kolmogorov, M., Parada, A.E., Weng, Z., Sidow, A., Dekas, A.E., Batzoglou, S., and Bhatt, A.S. (2018). High-quality genome sequences of uncultured microbes by assembly of read clouds. Nat. Biotechnol.

Bolourian, A., and Mojtahedi, Z. (2018). Streptomyces, shared microbiome member of soil and gut, as “old friends” against colon cancer. FEMS Microbiology Ecology *94*.

Bowers, R.M., Kyrpides, N.C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T.B.K., Schulz, F., Jarett, J., Rivers, A.R., Eloe-Fadrosh, E.A., et al. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat. Biotechnol. *35*, 725–731.

Brewster, R., Tamburini, F.B., Asiimwe, E., Oduaran, O., Hazelhurst, S., and Bhatt, A.S. (2019). Surveying Gut Microbiome Research in Africans: Toward Improved Diversity and Representation. Trends Microbiol.

Brito, I.L., Gurry, T., Zhao, S., Huang, K., Young, S.K., Shea, T.P., Naisilisili, W., Jenkins, A.P., Jupiter, S.D., Gevers, D., et al. (2019). Transmission of human-associated microbiota along family and social networks. Nat Microbiol *4*, 964–971.

Brown, C.T., and Irber, L. (2016). sourmash: a library for MinHash sketching of DNA. JOSS *1*, 27.

Campbell, T.P., Sun, X., Patel, V.H., Sanz, C., Morgan, D., and Dantas, G. (2020). The microbiome and resistome of chimpanzees, gorillas, and humans across host lifestyle and geography. ISME J. *14*, 1584–1599.

Ciabattini, A., Olivieri, R., Lazzeri, E., and Medaglini, D. (2019). Role of the Microbiota in the Modulation of Vaccine Immune Responses. Front. Microbiol. *10*, 1305.

Cole, J.R., Wang, Q., Fish, J.A., Chai, B., McGarrell, D.M., Sun, Y., Brown, C.T., Porras-Alfaro, A., Kuske, C.R., and Tiedje, J.M. (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res. *42*, D633–D642.

Collinson, M.A., White, M.J., Bocquier, P., McGarvey, S.T., Afolabi, S.A., Clark, S.J., Kahn, K., and Tollman, S.M. (2014). Migration and the epidemiological transition: insights from the Agincourt sub-district of northeast South Africa. Glob. Health Action *7*, 23514.

Crusoe, M.R., Alameldin, H.F., Awad, S., Boucher, E., Caldwell, A., Cartwright, R., Charbonneau, A., Constantinides, B., Edvenson, G., Fay, S., et al. (2015). The khmer software package: enabling efficient nucleotide sequence analysis. F1000Res. *4*, 900.

de la Cuesta-Zuluaga, J., Corrales-Agudelo, V., Velásquez-Mejía, E.P., Carmona, J.A., Abad, J.M., and Escobar, J.S. (2018). Gut microbiota is associated with obesity and cardiometabolic disease in a population in the midst of Westernization. Sci. Rep. *8*, 11356.

De Filippo, C., Cavalieri, D., Di Paola, M., Ramazzotti, M., Poullet, J.B., Massart, S., Collini, S., Pieraccini, G., and Lionetti, P. (2010). Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. Proc. Natl. Acad. Sci. U. S. A. *107*, 14691–14696.

Di Rienzi, S.C., Sharon, I., Wrighton, K.C., Koren, O., Hug, L.A., Thomas, B.C., Goodrich, J.K., Bell, J.T., Spector, T.D., Banfield, J.F., et al. (2013). The human gut and groundwater harbor non-photosynthetic bacteria belonging to a new candidate phylum sibling to Cyanobacteria. Elife *2*, e01102.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797.

Edwards, R.A., Vega, A.A., Norman, H.M., Ohaeri, M., Levi, K., Dinsdale, E.A., Cinek, O., Aziz, R.K., McNair, K., Barr, J.J., et al. (2019). Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. Nat Microbiol *4*, 1727–1736.

Fragiadakis, G.K., Smits, S.A., Sonnenburg, E.D., Van Treuren, W., Reid, G., Knight, R., Manjurano, A., Changalucha, J., Dominguez-Bello, M.G., Leach, J., et al. (2018). Links between environment, diet, and the hunter-gatherer microbiome. Gut Microbes *10*, 216–227.

Gentleman, R., Carey, V., Huber, W., and Hahne, F. (2019). genefilter: genefilter: methods for filtering genes from high-throughput experiments.

Ginsburg, C., Collinson, M.A., Iturralde, D., van Tonder, L., Gómez-Olivé, F.X., Kahn, K., and Tollman, S. (2016). Migration and Settlement Change in South Africa: Triangulating Census 2011 with Longitudinal Data from the Agincourt Health and Demographic Surveillance System in the Rural North-east. South. Afr. J. Demogr. *17*, 133–198.

Gonçalves da Silva, A. (2017). harrietr: Wrangle Phylogenetic Distance Matrices and Other Utilities.

Gorvitovskaia, A., Holmes, S.P., and Huse, S.M. (2016). Interpreting Prevotella and Bacteroides as biomarkers of diet and lifestyle. Microbiome *4*, 15.

Griffiths, J.A., and Mazmanian, S.K. (2018). Emerging evidence linking the gut microbiome to neurologic disorders. Genome Medicine *10*.

Guerin, E., Shkoporov, A., Stockdale, S.R., Clooney, A.G., Ryan, F.J., Sutton, T.D.S., Draper, L.A., Gonzalez-Tortuero, E., Ross, R.P., and Hill, C. (2018). Biology and Taxonomy of crAss-like Bacteriophages, the Most Abundant Virus in the Human Gut. Cell Host Microbe.

Gupta, V.K., Paul, S., and Dutta, C. (2017). Geography, Ethnicity or Subsistence-Specific Variations in Human Microbiome Composition and Diversity. Front. Microbiol. *8*, 1162.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics *29*, 1072–1075.

Hagan, T., Cortese, M., Rouphael, N., Boudreau, C., Linde, C., Maddur, M.S., Das, J., Wang, H., Guthmiller, J., Zheng, N.-Y., et al. (2019). Antibiotics-Driven Gut Microbiome Perturbation Alters Immunity to Vaccines in Humans. Cell *178*, 1313–1328.e13.

Han, C., Gronow, S., Teshima, H., Lapidus, A., Nolan, M., Lucas, S., Hammon, N., Deshpande, S., Cheng, J.-F., Zeytun, A., et al. (2011). Complete genome sequence of Treponema succinifaciens type strain (6091). Stand. Genomic Sci. *4*, 361–370.

Hansen, M.E.B., Rubel, M.A., Bailey, A.G., Ranciaro, A., Thompson, S.R., Campbell, M.C., Beggs, W., Dave, J.R., Mokone, G.G., Mpoloka, S.W., et al. (2019). Population structure of human gut bacteria in a diverse cohort from rural Tanzania and Botswana. Genome Biol. *20*, 16.

Helmink, B.A., Wadud Khan, M.A., Hermann, A., Gopalakrishnan, V., and Wargo, J.A. (2019). The microbiome, cancer, and cancer therapy. Nature Medicine *25*, 377–388.

Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. Nature *486*, 207–214.

Hunt, M., Silva, N.D., Otto, T.D., Parkhill, J., Keane, J.A., and Harris, S.R. (2015). Circlator: automated circularization of genome assemblies using long sequencing reads. Genome Biol. *16*, 294.

Jha, A.R., Davenport, E.R., Gautam, Y., Bhandari, D., Tandukar, S., Ng, K.M., Fragiadakis, G.K., Holmes, S., Gautam, G.P., Leach, J., et al. (2018). Gut microbiome transition across a lifestyle gradient in Himalaya. PLoS Biol. *16*, e2005396.

Kabudula, C.W., Houle, B., Collinson, M.A., Kahn, K., Gómez-Olivé, F.X., Clark, S.J., and Tollman, S. (2017a). Progression of the epidemiological transition in a rural South African setting: findings from population surveillance in Agincourt, 1993--2013. BMC Public Health *17*, 424.

Kabudula, C.W., Houle, B., Collinson, M.A., Kahn, K., Gómez-Olivé, F.X., Tollman, S., and Clark, S.J. (2017b). Socioeconomic differences in mortality in the antiretroviral therapy era in Agincourt, rural South Africa, 2001-13: a population surveillance analysis. Lancet Glob Health *5*, e924–e935.

Kang, D.D., Froula, J., Egan, R., and Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ *3*, e1165.

Kassambara, A. (2020). ggpubr: “ggplot2” Based Publication Ready Plots.

Koslicki, D., and Falush, D. (2016). MetaPalette: a -mer Painting Approach for Metagenomic Taxonomic Profiling and Quantification of Novel Strain Variation. mSystems *1*.

Krueger, F. Trim Galore!

Laslett, D., and Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. *32*, 11–16.

Leung, J.M., Graham, A.L., and Knowles, S.C.L. (2018). Parasite-Microbiota Interactions With the Vertebrate Gut: Synthesis Through an Ecological Lens. Front. Microbiol. *9*, 843.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics.

Li, D., Luo, R., Liu, C.-M., Leung, C.-M., Ting, H.-F., Sadakane, K., Yamashita, H., and Lam, T.-W. (2016). MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods *102*, 3–11.

Lim, S.S., Vos, T., Flaxman, A.D., Danaei, G., Shibuya, K., Adair-Rohani, H., Amann, M., Anderson, H.R., Andrews, K.G., Aryee, M., et al. (2012). A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet *380*, 2224–2260.

Lin, Y., Yuan, J., Kolmogorov, M., Shen, M.W., Chaisson, M., and Pevzner, P.A. (2016). Assembly of long error-prone reads using de Bruijn graphs. Proc. Natl. Acad. Sci. U. S. A. *113*, E8396–E8405.

Lokmer, A., Cian, A., Froment, A., Gantois, N., Viscogliosi, E., Chabé, M., and Ségurel, L. (2019). Use of shotgun metagenomics for the identification of protozoa in the gut microbiota of healthy individuals from worldwide populations with various industrialization levels. PLoS One *14*, e0211139.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Lu, J., Breitwieser, F.P., Thielen, P., and Salzberg, S.L. (2017). Bracken: estimating species abundance in metagenomics data. PeerJ Comput. Sci. *3*, e104.

Maier, L., and Typas, A. (2017). Systematically investigating the impact of medication on the gut microbiome. Curr. Opin. Microbiol. *39*, 128–135.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal *17*, 10–12.

Martínez, I., Stegen, J.C., Maldonado-Gómez, M.X., Eren, A.M., Siba, P.M., Greenhill, A.R., and Walter, J. (2015). The gut microbiota of rural papua new guineans: composition, diversity patterns, and ecological processes. Cell Rep. *11*, 527–538.

Moss, E.L., Maghini, D.G., and Bhatt, A.S. (2020). Complete, closed bacterial genomes from microbiomes using nanopore sequencing. Nat. Biotechnol.

Nayfach, S., Shi, Z.J., Seshadri, R., Pollard, K.S., and Kyrpides, N.C. (2019). New insights from uncultivated genomes of the global human gut microbiome. Nature *568*, 505–510.

NCD Risk Factor Collaboration (NCD-RisC) – Africa Working Group (2017). Trends in obesity and diabetes across Africa from 1980 to 2014: an analysis of pooled population-based studies. Int. J. Epidemiol. *46*, 1421–1432.

Obregon-Tito, A.J., Tito, R.Y., Metcalf, J., Sankaranarayanan, K., Clemente, J.C., Ursell, L.K., Zech Xu, Z., Van Treuren, W., Knight, R., Gaffney, P.M., et al. (2015). Subsistence strategies in traditional societies distinguish gut microbiomes. Nat. Commun. *6*, 6505.

Oduaran, O.H., Tamburini, F.B., Sahibdeen, V., Brewster, R., Gómez-Olivé, F.X., Kahn, K., Norris, S.A., Tollman, S.M., Twine, R., Wade, A.N., et al. (2020). Gut Microbiome Profiling of a Rural and Urban South African Cohort Reveals Biomarkers of a Population in Lifestyle Transition. Biorxiv.

Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., et al. (2019). vegan: Community Ecology Package.

Olm, M.R., Brown, C.T., Brooks, B., and Banfield, J.F. (2017). dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J. *11*, 2864–2868.

Ou, J., Carbonero, F., Zoetendal, E.G., DeLany, J.P., Wang, M., Newton, K., Gaskins, H.R., and O'Keefe, S.J.D. (2013). Diet, microbiota, and microbial metabolites in colon cancer risk in rural Africans and African Americans. Am. J. Clin. Nutr. *98*, 111–120.

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. *25*, 1043–1055.

Parks, D.H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B.J., Evans, P.N., Hugenholtz, P., and Tyson, G.W. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol *2*, 1533–1542.

Parks, D.H., Chuvochina, M., Chaumeil, P.-A., Rinke, C., Mussig, A.J., and Hugenholtz, P. (2020). A complete domain-to-species taxonomy for Bacteria and Archaea. Nat. Biotechnol. *38*, 1079–1086.

Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., et al. (2019). Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. Cell *176*, 649–662.e20.

Paulson, J.N., Stine, O.C., Bravo, H.C., and Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. Nat. Methods *10*, 1200–1202.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One *5*, e9490.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. Nature *464*, 59–65.

Rampelli, S., Schnorr, S.L., Consolandi, C., Turroni, S., Severgnini, M., Peano, C., Brigidi, P., Crittenden, A.N., Henry, A.G., and Candela, M. (2015). Metagenome Sequencing of the Hadza Hunter-Gatherer Gut Microbiota. Curr. Biol. *25*, 1682–1693.

Ramsay, M., Crowther, N., Tambo, E., Agongo, G., Baloyi, V., Dikotope, S., Gómez-Olivé, X., Jaff, N., Sorgho, H., Wagner, R., et al. (2016). The AWI-Gen Collaborative Centre: Understanding the interplay between Genomic and Environmental Risk Factors for Cardiometabolic Diseases in sub-Saharan Africa. Global Health, Epidemiology and Genomics.

R Core Team (2019). R: A Language and Environment for Statistical Computing.

Richter, L., Norris, S., Pettifor, J., Yach, D., and Cameron, N. (2007). Cohort Profile: Mandela's children: the 1990 Birth to Twenty study in South Africa. Int. J. Epidemiol. *36*, 504–511.

Rothschild, D., Weissbrod, O., Barkan, E., Kurilshikov, A., Korem, T., Zeevi, D., Costea, P.I., Godneva, A., Kalka, I.N., Bar, N., et al. (2018). Environment dominates over host genetics in shaping human gut microbiota. Nature *555*, 210–215.

Santosa, A., and Byass, P. (2016). Diverse Empirical Evidence on Epidemiological Transition in Low-

and Middle-Income Countries: Population-Based Findings from INDEPTH Network Data. PLoS One *11*, e0155753.

Sato, M.P., Ogura, Y., Nakamura, K., Nishida, R., Gotoh, Y., Hayashi, M., Hisatsune, J., Sugai, M., Takehiko, I., and Hayashi, T. (2019). Comparison of the sequencing bias of currently available library preparation kits for Illumina sequencing of bacterial genomes and metagenomes. DNA Res. *26*, 391–398.

Scher, J.U., Sczesnak, A., Longman, R.S., Segata, N., Ubeda, C., Bielski, C., Rostron, T., Cerundolo, V., Pamer, E.G., Abramson, S.B., et al. (2013). Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis. Elife *2*, e01202.

Schnorr, S.L., Candela, M., Rampelli, S., Centanni, M., Consolandi, C., Basaglia, G., Turroni, S., Biagi, E., Peano, C., Severgnini, M., et al. (2014). Gut microbiome of the Hadza hunter-gatherers. Nat. Commun. *5*, 3654.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics *30*, 2068–2069.

Slowikowski, K. (2020). ggrepel: Automatically Position Non-Overlapping Text Labels with "ggplot2."

Smits, S.A., Leach, J., Sonnenburg, E.D., Gonzalez, C.G., Lichtman, J.S., Reid, G., Knight, R., Manjurano, A., Changalucha, J., Elias, J.E., et al. (2017). Seasonal cycling in the gut microbiome of the Hadza hunter-gatherers of Tanzania. Science *357*, 802–806.

Sonnenburg, E.D., and Sonnenburg, J.L. (2019). The ancestral and industrialized gut microbiota and implications for human health. Nat. Rev. Microbiol. *17*, 383–390.

Sonnenburg, J., and Sonnenburg, E. (2018). A Microbiota Assimilation. Cell Metab. *28*, 675–677.

Soo, R.M., Skennerton, C.T., Sekiguchi, Y., Imelfort, M., Paech, S.J., Dennis, P.G., Steen, J.A., Parks, D.H., Tyson, G.W., and Hugenholtz, P. (2014). An expanded genomic representation of the phylum cyanobacteria. Genome Biol. Evol. *6*, 1031–1045.

Statistics South Africa (2012). Census 2011 Statistical Release.

Stewart, R.D., Auffret, M.D., Warr, A., Walker, A.W., Roehe, R., and Watson, M. (2019). Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. Nat. Biotechnol. *37*, 953–961.

Tett, A., Huang, K.D., Asnicar, F., Fehlner-Peach, H., Pasolli, E., Karcher, N., Armanini, F., Manghi, P., Bonham, K., Zolfo, M., et al. (2019). The Prevotella copri Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations. Cell Host Microbe *26*, 666–679.e7.

Trönnberg, L., Hawksworth, D., Hansen, A., Archer, C., and Stenström, T.A. (2010). Household-based prevalence of helminths and parasitic protozoa in rural KwaZulu-Natal, South Africa, assessed from faecal vault sampling. Trans. R. Soc. Trop. Med. Hyg. *104*, 646–652.

Turnbaugh, P.J., Hamady, M., Yatsunenko, T., Cantarel, B.L., Duncan, A., Ley, R.E., Sogin, M.L., Jones, W.J., Roe, B.A., Affourtit, J.P., et al. (2009). A core gut microbiome in obese and lean twins. Nature *457*, 480–484.

Tyler, A.D., Mataseje, L., Urfano, C.J., Schmidt, L., Antonation, K.S., Mulvey, M.R., and Corbett, C.R. (2018). Evaluation of Oxford Nanopore's MinION Sequencing Device for Microbial Whole Genome Sequencing Applications. Sci. Rep. *8*, 10931.

Vangay, P., Johnson, A.J., Ward, T.L., Al-Ghalith, G.A., Shields-Cutler, R.R., Hillmann, B.M., Lucas,

S.K., Beura, L.K., Thompson, E.A., Till, L.M., et al. (2018). US Immigration Westernizes the Human Gut Microbiome. Cell *175*, 962–972.e10.

Vaser, R., Sović, I., Nagarajan, N., and Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. *27*, 737–746.

Venables, W.N., and Ripley, B.D. (2002). Modern Applied Statistics with S.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One *9*, e112963.

Warnes, G.R., Bolker, B., and Lumley, T. (2020). gtools: Various R Programming Tools.

Wickham, H. (2007). Reshaping Data with the reshape Package. J. Stat. Softw. *21*, 1–20.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis.

Wickham, H., François, R., Henry, L., and Müller, K. (2020). dplyr: A Grammar of Data Manipulation.

Wilke, C.O. (2019). cowplot: Streamlined Plot Theme and Plot Annotations for "ggplot2."

Wood, D.E., and Salzberg, S.L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. *15*, R46.

Yatsunenko, T., Rey, F.E., Manary, M.J., Trehan, I., Dominguez-Bello, M.G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R.N., Anokhin, A.P., et al. (2012). Human gut microbiome viewed across age and geography. Nature *486*, 222–227.