Short- and long-read metagenomics of urban and rural 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], F. 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: Scott.Hazelhurst@wits.ac.za, asbhatt@stanford.edu
[+]Lead contact

Abstract

Human gut microbiome research focuses on populations living in high-income countries or on the other end of the spectrum, namely non-urban agriculturalist and hunter-gatherer societies. The scarcity of research between these extremes limits our understanding of how the gut microbiota relates to health and disease in the majority of the world's population. 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=118) or urban Soweto (n=51) and find that these microbiomes are taxonomically intermediate between those of individuals living in high-income countries and traditional communities. We demonstrate that reference collections are incomplete for characterization of microbiomes of individuals living outside high-income countries, resulting in artificially low species-level beta diversity measurements. To improve reference databases, we generated complete genomes of undescribed taxa, including *Treponema*, Lentisphaerae, and *Succinatimonas* species. Our results suggest that the gut microbiome in South African populations do not exist along a simple “western-nonwestern” axis and that these populations contain microbial diversity that remains to be described.

Introduction

Comprehensive characterization of the full diversity of the healthy human gut microbiota is essential to contextualize studies of the microbiome related to diet, lifestyle, and disease. To date, substantial resources have been invested in describing the microbiome of individuals living in the global industrialized "west" (United States, northern and western Europe; also sometimes referred to as the "Global North"), including efforts by large consortia such as the Human Microbiome Project[1] and MetaHIT[2]. Though these projects have yielded valuable descriptions of human gut microbial ecology, they survey only a small portion of the world's citizens at the extreme of industrialized, urbanized lifestyle. It is unclear to what extent these results are generalizable to non-western and non-industrialized populations across the globe.

At the other extreme, a smaller number of studies have characterized the gut microbiome composition of individuals practicing traditional lifestyles[3,4], including communities in Venezuela and Malawi[5], hunter-gatherer communities in Tanzania[6–9], non-industrialized populations in Tanzania and Botswana[10], and agriculturalists in Peru[11] and remote Madagascar[12]. However, these cohorts are not representative of how most of the world lives either. Many of the world's communities lead lifestyles between the extremes of an urbanized, industrialized and relatively high-income lifestyle and traditional subsistence practices. It is a scientific and ethical imperative to include these diverse populations in biomedical research, yet dismayingly many of these intermediate groups are underrepresented in or absent from the published microbiome literature.

This major gap in our knowledge of the human gut microbiome leaves the biomedical research community ill-poised to relate microbiome composition to human health and disease across the breadth of the world's population. Worldwide, many communities are currently undergoing a transition of diet and lifestyle, characterized by increased access to processed foods, diets rich in animal fats and simple carbohydrates, and more sedentary lifestyles[13]. This has corresponded with an epidemiological transition in which the burden of disease is shifting from predominantly infectious diseases to an increasing incidence of noncommunicable diseases (NCDs) like obesity and diabetes[14]. The microbiome has been implicated in various NCDs[15–17]

79 and may mediate the efficacy of medical interventions including vaccines[18,19], but we
80 cannot evaluate the generalizability of these findings without establishing baseline
81 microbiome characteristics of communities that practice diverse lifestyles and by
82 extension, harbor diverse microbiota. These understudied populations, which are more
83 representative of the majority of the world's population, offer a unique opportunity to
84 examine the relationship between lifestyle (including diet), disease, and gut microbiome
85 composition, and to discover novel microbial genomic content that may associate with
86 or drive disease biology.

87 Some previous studies have probed the relationship between lifestyle and
88 microbiome composition in transitional communities[3,20–22]. However, substantial gaps
89 remain in our description of the microbiome in these populations. In particular,
90 knowledge of the gut microbiota within the African continent is sparse. Of the 64
91 studies surveying the gut microbiome of individuals living within Africa as of January
92 2021 (Supplementary Table 1) only 25 of the 54 countries (46%) on the continent are
93 represented. Of these studies, 34 of 64 (53%) have focused entirely on children or
94 infants, whose disease risk profile and gut microbiome composition can vary
95 considerably from adults[5,23]. Additionally, 52 of 64 (81%) of studies of the gut
96 microbiome in Africans employed 16S rRNA gene sequencing or qPCR, techniques
97 which amplify only a small portion of the genome and therefore lack genomic
98 resolution to describe species or strains which may share a 16S rRNA sequence but
99 differ in gene content or genome structure. To our knowledge, only nine published
100 studies to date have used shotgun metagenomics to describe the gut microbiome of
101 adults living in Africa. Eight of these studies described the bacterial microbiome[6,7,12,24–
102 28], while one[29] exclusively described the viral metagenome.

103 To address this major knowledge gap, we designed and performed the first
104 research study applying short- and long-read DNA sequencing to study the gut
105 microbiomes of South African individuals for whom 16S rRNA gene sequence data has
106 recently been reported[30]. South Africa is a prime example of a country undergoing
107 rapid lifestyle and epidemiological transition. With the exception of the HIV/AIDS
108 epidemic in the mid-1990s to the mid-2000s, over the past three decades South Africa
109 has experienced a steadily decreasing mortality rate from infectious disease and an

110 increase in NCD[31,32]. Concomitantly, increasingly sedentary lifestyles and changes in
111 dietary habits, including access to calorie-dense processed foods, contribute to a
112 higher prevalence of obesity in many regions of South Africa[32], a trend which
113 disproportionately affects women[33,34].

114 This study presents the largest shotgun metagenomic dataset of African adults
115 in the published literature to date. In this work, we describe microbial community-scale
116 similarities between urban and rural communities in South Africa, as well as distinct
117 hallmark taxa that distinguish each community. Additionally, we place South African
118 populations in context with microbiome data from other populations from countries
119 around the world, revealing the transitional nature of gut microbiome composition in
120 the South African cohorts. We demonstrate that metagenomic assembly of short reads
121 yields novel strain and species draft genomes. Finally, we apply Oxford Nanopore
122 long-read sequencing to samples from the rural cohort and generate complete and
123 near-complete genomes. These include genomes of species that are exclusive to, or
124 more prevalent in, traditional populations, including *Treponema* and *Prevotella* species.
125 As long-read sequencing enables more uniform coverage of AT-rich regions compared
126 to short-read sequencing with transposase-based library preparation, we also generate
127 complete metagenome-assembled AT-rich genomes from less well-described gut
128 microbes including species in the phylum *Melainabacteria*, the class *Mollicutes*, and
129 the genus *Mycoplasma*.

130 Taken together, the results herein offer a more detailed description of gut
131 microbiome composition in understudied transitioning populations, and present
132 complete and contiguous reference genomes that will enable further studies of gut
133 microbiota in nonwestern populations. Importantly, this study was developed with an
134 ethical commitment to engaging both rural and urban community members to ensure
135 that the research was conducted equitably (additional details in Supplemental
136 Information). This work underscores the critical need to broaden the scope of human
137 gut microbiome research and include understudied, nonwestern populations to
138 improve the relevance and accuracy of microbiome discoveries to broader populations.

Results

Cohorts and sample collection

We enrolled 190 women aged between 40-72, living in rural villages in the Bushbuckridge Municipality (24.82°S, 31.26°E, n=132) and urban Soweto, Johannesburg (26.25°S, 27.85°E, n=58) and collected a one-time stool sample, as well as point of care blood glucose and blood pressure measurements and a rapid HIV test. As HIV status and exposure to antiretroviral medications can alter the microbiome and potentially confound analyses, only samples from HIV-negative individuals were analyzed further (n=118 Bushbuckridge, n=51 Soweto). Participants spanned a range of BMI from healthy to overweight; the most common comorbidity reported was hypertension, and many patients reported taking anti-hypertensive medication (18 of 118 (15%) in Bushbuckridge, 15 of 51 (29%) in Soweto) (Table 1, Supplementary Table 2). Additional medications are summarized in Supplementary Table 2. We extracted DNA from each stool sample and conducted 150 base pair (bp) paired-end sequencing on the Illumina HiSeq 4000 platform. A median of 34.6 million (M) raw reads were generated per sample (range 11.4-100 M), and a median of 14.9 M reads (range 4.2-33.3 M) resulted after preprocessing including de-duplication, trimming, and human read removal (Supplementary Table 3).

Gut microbial composition

We taxonomically classified sequencing reads against a comprehensive custom reference database containing all microbial genomes in RefSeq and GenBank at "scaffold" quality or better as of January 2020 (177,626 genomes total). Concordant with observations from 16S rRNA gene sequencing of the same samples[30], we find that *Prevotella*, *Bacteroides,* and *Faecalibacterium* are the most abundant genera in most individuals across both study sites (Figure 1A, Supplementary Fig. 1, Supplementary Table 4; species-level classifications in Supplementary Table 5). Additionally, in many individuals we observe taxa that are uncommon in western microbiomes, including members of the VANISH (Volatile and/or Associated Negatively with Industrialized Societies of Humans) taxa (families *Prevotellaceae*, *Succinovibrionaceae*,

170 *Paraprevotellaceae*, and *Spirochaetaceae*) such as *Prevotella, Treponema*, and
171 *Succinatimonas*, which are higher in relative abundance in communities practicing
172 traditional lifestyles compared to western industrialized populations[8,35] (Figure 1B,
173 Supplementary Table 4). The mean relative abundance of each VANISH genus is higher
174 in Bushbuckridge than Soweto, though the difference is not statistically significant for
175 *Paraprevotella* or *Sediminispirochaeta* (Figure 1B, two-sided Wilcoxon rank sum test).
176 Within the Bushbuckridge cohort, we observe a bimodal distribution of the genera
177 *Succinatimonas*, *Succinivibrio*, and *Treponema* (Supplementary Fig. 2A). While we do
178 not identify any clinical or demographic features that associate with this distribution,
179 we observe that VANISH taxa are weakly positively correlated with one another in
180 metagenomes from both Bushbuckridge and Soweto (Supplementary Fig. 2B-C).
181 Intriguingly, we observed that an increased proportion of reads aligned to the
182 human genome during pre-processing in samples from Soweto compared to
183 Bushbuckridge (Supplementary Fig. 3, two-sided Wilcoxon rank sum test $p < 0.0001$).
184 This could potentially indicate higher inflammation and immune cell content or
185 sloughing of intestinal epithelial cells in the urban Soweto cohort compared to rural
186 Bushbuckridge.
187
188 ***Rural and urban microbiomes cluster distinctly in MDS***
189 We hypothesized that lifestyle differences of those residing in rural
190 Bushbuckridge versus urban Soweto might be associated with demonstrable
191 differences in gut microbiome composition. Bushbuckridge and Soweto differ markedly
192 in their population density (53 and 6,357 persons per km^2 respectively as of the 2011
193 census) as well as in lifestyle variables including the prevalence of flush toilets (6.8 vs
194 91.6% of dwellings) and piped water (11.9 vs 55% of dwellings) (additional site
195 demographic information in Supplementary Table 6)[36]. Soweto is highly urbanized and
196 has been so for several decades, while Bushbuckridge is classified as a rural
197 community, although it is undergoing rapid epidemiological transition[37,38].
198 Bushbuckridge also has circular rural/urban migrancy typified by some (mostly male)
199 members of a rural community working and living for extended periods in urban areas,
200 while keeping their permanent rural home[39]. Although our participants all live in

201 Bushbuckridge, this migrancy in the community contributes to making the boundary
202 between rural and urban lifestyles more fluid. Comparing the two study populations at
203 the community level, we find that samples from the two sites have distinct centroids
204 (PERMANOVA $p < 0.001$, $R^2 = 0.037$) but overlap (Figure 2A), though we note that the
205 dispersion of the Soweto samples is greater than that of the Bushbuckridge samples
206 (PERMDISP2 $p < 0.001$). Across the study population we observe a gradient of
207 *Bacteroides* and *Prevotella* relative abundance (Supplementary Fig. 4). This may be the
208 result of differences in diet across the study population at both sites, as *Bacteroides*
209 has been proposed as a biomarker of westernized lifestyles while *Prevotella* has been
210 proposed as a biomarker of nonwestern lifestyles[5,40,41].
211 To determine if medication usage was associated with gut microbiome
212 composition, we included each participant's self-reported concomitant medications
213 (summarized in Supplementary Table 2) to re-visualize the microbiome composition of
214 samples in MDS by class of medication (Supplementary Fig. 5A,B). We find that self-
215 reported medication is not significantly correlated with community composition in this
216 cohort after multiple hypothesis correction (PERMANOVA $q > 0.05$, Supplementary Fig.
217 5C), though two drug classes are nominally significant before controlling the false
218 discovery rate: proton pump inhibitors (PPIs) ($p = 0.036$) and anti-hyperglycemics ($p =$
219 0.041). We note that both drug classes have previously been found to associate with
220 changes in gut microbiome composition[42–44]: as only two participants self-report taking
221 PPIs at the time of sampling, additional data are required to evaluate whether PPIs
222 associate with microbiome composition in these South African populations.
223
224 ***Rural and urban microbiomes differ in Shannon diversity and species***
225 ***composition***
226 Gut microbiome alpha diversity of individuals living traditional lifestyles has been
227 reported to be higher than those living western lifestyles[9,11,40]. In keeping with this
228 general trend, we find that alpha diversity (Shannon) is significantly higher in individuals
229 living in rural Bushbuckridge than urban Soweto (Figure 2B; two-sided Wilcoxon rank
230 sum test, $p < 0.01$). Using DESeq2 to identify microbial genera that are differentially
231 abundant across study sites, we find that genera including *Bacteroides*,

232 *Bifidobacterium,* and *Streptococcus* are more abundant in individuals living in Soweto
233 (Figure 2C, Supplementary Table 7, species shown in Supplementary Fig. 6).
234 Interestingly, we find microbial genera enriched in gut microbiomes of individuals living
235 in Bushbuckridge that are common to both the environment and the gut, including
236 *Streptomyces* and *Paenibacillus* (Supplementary Table 7). Typically a soil-associated
237 organism, *Streptomyces* encode a variety of biosynthetic gene clusters and can
238 produce numerous immunomodulatory and anti-inflammatory compounds such as
239 rapamycin and tacrolimus, and it has been suggested that decreased exposure to
240 *Streptomyces* is associated with increased incidence of inflammatory disease and
241 colon cancer in western populations[45]. In addition, we find enrichment of genera in
242 Bushbuckridge that have been previously associated with nonwestern microbiomes
243 including *Succinatimonas*, a relatively poorly-described bacterial genus with only one
244 type species, and unclassified species of the phylum Elusimicrobia, which has been
245 detected in the gut microbiome of rural Malagasy[12]. Additionally, Bushbuckridge
246 samples are enriched for Cyanobacteria as well as Candidatus Melainabacter, a
247 phylum closely related to Cyanobacteria that in limited studies has been described to
248 inhabit the human gut[46,47].
249 In terms of the non-bacterial microbiome, we identify the bacteriophage
250 crAssphage and related crAss-like phages[48], which have recently been described as
251 prevalent constituents of the gut microbiome globally[49], in 32 of 51 participants (63%)
252 in Soweto and 88 of 118 (75%) in Bushbuckridge (difference in prevalence between
253 cohorts not significant, $p = 0.14$ Fisher's exact test) using 650 sequence reads or
254 roughly 1X coverage of the 97 kb genome as a threshold for binary categorization of
255 crAss-like phage presence or absence. Prototypical crAssphage has been
256 hypothesized to infect *Bacteroides* species and a crAss-like phage has been
257 demonstrated to infect *Bacteroides intestinalis*. Though crAss-like phages do not differ
258 between cohorts in terms of prevalence (presence/absence), we observe that
259 crAssphage clade Delta from Guerin *et al.*[48] is enriched in relative abundance in the gut
260 microbiome of individuals living in Bushbuckridge compared to Soweto, supporting
261 previous observations of geographic patterns of crAssphage clades (Figure 2C)[49].

262 Our custom reference database of GenBank genomes paired with the kraken2
263 classifier optimizes for sensitivity; thus, this approach was selected as the initial tool for
264 classification of the sequencing data given the genomic novelty anticipated in this
265 cohort. We note that broadly similar microbiome profiles are obtained using
266 MetaPhlAn3, a marker-gene based tool with high specificity, (Supplementary Fig. 7) as
267 well as classifications obtained using kraken2 and a publicly available build of the
268 Genome Taxonomy Database (GTDB) release 95[50,51]. Notably, we observe higher
269 Shannon diversity with the GTDB compared to both MetaPhlAn3 and our custom
270 database, likely due to the fact that clades containing a large amount of genomic
271 diversity (e.g. *Escherichia coli*) are split into separate clades in the GTDB.
272
273 ***Differences in functional potential of the gut microbiome between populations***
274 Recognizing that functional annotations are likely biased toward well-studied
275 organisms, we sought to identify differentially abundant functions in the gut
276 microbiome of participants in Bushbuckridge and Soweto.
277 We functionally profiled unassembled metagenomic reads to detect antibiotic
278 resistance genes in these communities. Tetracycline resistance genes (*tetW, tetQ,*
279 *tetO, tetX, tet32, tet40*) are broadly prevalent in both populations (Supplementary Fig.
280 8) as is the CfxA6 beta-lactamase. We find that Soweto and Bushbuckridge differ in
281 the distribution of relative abundance of 30 of 113 (27%) antibiotic resistance genes
282 (Supplementary Fig. 8). Several multidrug efflux pump components and regulators
283 (*mdtB, mdtC, mdtF, mdtG, mdtL, mdtP, CRP*) are enriched in participants in
284 Bushbuckridge, whereas genes including *SAT-4*, which is a plasmid-encoded
285 streptothricin resistance determinant, and *CblA-1*, which encodes a class A beta-
286 lactamase, are enriched in Soweto participants (Supplementary Fig. 8).
287 We additionally annotated MetaCyc pathway abundance using HUMAnN v3[52]
288 (Supplementary Table 8). We find 68 MetaCyc pathways that are differentially abundant
289 between Soweto and Bushbuckridge ($q < 0.05$) (Supplementary Fig. 9A). Some of
290 these pathways correspond clearly to observed taxonomic differences between study
291 sites, including enrichment of the *Bifidobacterium* shunt, a pathway for degradation of
292 hexose sugars into short chain fatty acids[53], in Soweto. Other differentially abundant

293 pathways include anaerobic degradation of 4-coumarate, a phenylpropanoid
294 compound produced by plants and by catabolism of the amino acid tyrosine[54].
295 Additionally, the superpathway of phenylethylamine degradation is enriched in
296 Bushbuckridge. Intriguingly, phenylethylamine is a central nervous system stimulant in
297 humans and increased abundance of phenylethylamine has been observed in Crohn's
298 disease patients[55]. Peptidoglycan biosynthesis V pathway, involved in microbial
299 resistance to beta-lactam antibiotics, is enriched in Soweto, consistent with results
300 from antibiotic resistome profiling.
301 In general, HUMAnN was only able to ascribe functions to taxonomy for a few
302 well-studied genera including *Escherichia* and *Klebsiella* (Supplementary Fig. 9B). We
303 hypothesize that this is due to gaps in reference genome collections as well as
304 dissimilarity between strains of species that are common to reference collections and
305 metagenomic data from this cohort.
306
307 ***No strong signals of interaction between human DNA variation and microbiome***
308 ***content detected***
309 All participants in this study were recruited based on their participation in the
310 first phase of the Africa Wits-INDEPTH partnership for Genomic Studies (AWI-Gen)
311 study, which evaluated genomic and environmental risk factors for cardiometabolic
312 disease in sub-Saharan African populations[56]. This study included human genome
313 profiling of all participants using the Human Heredity and Health in Africa (H3Africa)
314 single nucleotide polymorphism (SNP) array. While we have a very small sample size to
315 assess interaction between human genetic variation and microbiome population, our
316 study is one of the relatively few to characterize both human and microbiome DNA.
317 Therefore, we performed association tests between key microbiome genera
318 abundance levels and human SNPs. After correcting for multiple testing there were
319 only a few human genomic SNPs with borderline statistically significant association
320 with microbial genera abundance levels (Supplementary Table 9). These SNPs occur in
321 genomic regions with no obvious connection to the gut microbiome (see Methods,
322 Supplementary Information). Additionally, we observe that microbiome samples do not
323 cluster by self-reported ethnicity of the participant (Supplementary Fig. 10).

324

325 ***South African gut microbiomes share taxa with western and nonwestern***
326 ***populations yet harbor distinct features***

327 To place the microbiome composition of South African individuals in global
328 context with metagenomes from healthy adults living in other parts of the world, we
329 compared publicly available data from five cohorts (Figure 3A, Supplementary Table
330 10) comprising adult individuals living in the United States[1], northern Europe
331 (Sweden)[57], agriculturalists living in Burkina Faso[28] and rural Madagascar[12], and the
332 Hadza hunter-gatherers of Tanzania[7]. We grouped these datasets by lifestyle into the
333 general categories of "nonwestern" (Tanzania, Madagascar, Burkina Faso), "western"
334 (USA, Sweden), and South African (Bushbuckridge, Soweto). We note the caveat that
335 these samples were collected at different times using different approaches, and that
336 there is variation in DNA extraction, sequencing library preparation and sequencing, all
337 of which may contribute to variation between studies. Recognizing this limitation, we
338 observe that South African samples cluster between western and nonwestern
339 populations in MDS (Figure 3B) as expected, and that the first axis of MDS correlates
340 well with geography and lifestyle (Figure 3C). The relative abundance of
341 *Spirochaetaceae*, *Succinivibrionaceae*, *Bacteroidaceae, and Prevotellaceae* are most
342 strongly correlated with the first axis of MDS (Spearman's rho > 0.75): *Bacteroidaceae*
343 decreases with MDS 1 while *Spirochaetaceae*, *Succinivibrionaceae,* and *Prevotellaceae*
344 increase (Figure 3B). We observe a corresponding pattern of decreasing relative
345 abundance of other VANISH taxa across lifestyle and geography (Supplementary Fig.
346 11). These observations suggest that the gut microbiome of South African cohorts is to
347 some extent "intermediate" in composition when compared to cohorts at the extremes
348 of western and nonwestern lifestyle.

349 The two South African cohorts also have distinct differences from both
350 nonwestern and western populations, as evidenced by displacement along the second
351 axis of MDS (Figure 3B,C). To identify the taxa that drive this separation, we performed
352 statistical analysis using DESeq2 to identify microbial genera that differed significantly
353 in the South African cohort compared to both nonwestern and western categories (with
354 the same directionality of effect in each comparison, e.g. enriched in South Africans

355 compared to both western and nonwestern groups) (Supplementary Fig. 12). We
356 observe that taxa including *Lactobacillus*, *Lactococcus*, and *Eggerthella* are lower in
357 relative abundance in South Africans compared to both western and nonwestern
358 microbiomes. Conversely, *Klebsiella* and unclassified *Christensenellaceae* are enriched
359 in South Africans.
360
361 ***Within-species diversity across cohorts***
362 Having observed taxonomic differences at the species level between South
363 Africans and other global populations, as well as between Soweto and Bushbuckridge,
364 we hypothesized that strains of some species may differ between populations. We
365 annotated the pangenome of the top six most abundant species on average across our
366 cohorts and assessed whether pangenome content is significantly different between
367 study sites (Supplementary Fig. 13). Interestingly, we find that *F. prausnitzii, B.*
368 *vulgatus,* and *E. siraeum* indeed differ in pangenome content between Bushbuckridge
369 and Soweto. *Prevotella copri* strains exhibit visible heterogeneity, but a PERMANOVA
370 test is not significant after false discovery rate correction.
371
372 ***Decreased sequence classifiability in nonwestern populations***
373 Given previous observations that gut microbiome alpha diversity is higher in
374 individuals practicing traditional lifestyles[3,6,58] and that immigration from Southeast Asia
375 to the United States is associated with a decrease in gut microbial alpha diversity[13], we
376 hypothesized that alpha diversity would be higher in nonwestern populations, including
377 South Africans, compared to western populations. We observe that Shannon diversity
378 of the Tanzanian hunter-gatherer cohort is uniformly higher than all other populations
379 (Figure 3D; $q < 0.05$ for all pairwise comparisons; FDR-adjusted two-sided Wilcoxon
380 rank sum test) and that alpha diversity is lower in individuals living in the United States
381 compared to all other cohorts (Figure 3D; $q < 0.0001$ for all pairwise comparisons;
382 FDR-adjusted two-sided Wilcoxon rank sum test). Surprisingly, we observe
383 comparable Shannon diversity between Madagascar and Sweden ($q > 0.05$, two-sided
384 Wilcoxon rank sum test). However, this could be an artifact of incomplete
385 representation of diverse microbes in existing reference collections.

386 Existing reference collections are known to be limited in their ability to classify
387 metagenomic sequences from nonwestern gut microbiomes[12,59], and we observe low
388 sequence classifiability in nonwestern populations (Figure 4A). Therefore, we sought
389 orthogonal validation of our observation that South African microbiomes represent a
390 transitional state between traditional and western microbiomes and employed a
391 reference-independent method to evaluate the nucleotide composition of sequence
392 data from each metagenome. We used the sourmash workflow[60] to compare
393 nucleotide *k*-mer composition of sequencing reads in each sample and performed
394 ordination based on angular distance, which accounts for *k*-mer abundance. Using a *k*-
395 mer length of 31 (*k*-mer similarity at *k*=31 correlates with species-level similarity[61]), we
396 observe clustering reminiscent of the species ordination plot shown in Fig. 3, further
397 supporting the hypothesis that South African microbiomes are transitional (Figure 4B).
398 Previous studies have reported a pattern of higher alpha diversity but lower beta
399 diversity in nonwestern populations compared to western populations[9,62].
400 Hypothesizing that alpha and beta diversity may be underestimated for populations
401 whose gut microbes are not well-represented in reference collections, we compared
402 beta diversity (distributions of within-cohort pairwise distances) calculated via species
403 Bray-Curtis dissimilarity as well as nucleotide *k*-mer angular distance (Figure 4C-E). Of
404 note, beta diversity is highest in Soweto irrespective of distance measure (Figure 4C).
405 Intriguingly, in some cases we observe that the relationship of distributions of pairwise
406 distance values changes depending on whether species or nucleotide *k*-mers are
407 considered. For instance, considering only species content, Bushbuckridge has less
408 beta diversity than Sweden, but this pattern is reversed when considering nucleotide *k*-
409 mer content (Figure 4D). Further, the same observation is true for the relationship
410 between Madagascar and the United States (Figure 4E). Additionally, we compared
411 species and nucleotide beta diversity within each population using Jaccard distance,
412 which is computed based on shared and distinct features irrespective of abundance. In
413 nucleotide *k*-mer space, all nonwestern populations have greater beta diversity than
414 each western population (Supplementary Fig. 14), though this is not the case when
415 species annotations are considered. This indicates that gut microbiomes in these

416 nonwestern cohorts have a longer “tail” of lowly abundant organisms which differ
417 between individuals.
418 These observations are critically important to our understanding of beta diversity
419 in the gut microbiome in western and nonwestern communities. In summary, we find
420 evidence to refute the existing dogma of an inverse relationship between alpha and
421 beta diversity, and note that in some cases this existing generalization represents an
422 artifact of limitations in reference databases used for sequence classification.
423
424 ***Improving reference collections via metagenomic assembly***
425 Classification of metagenomic sequencing reads can be improved by
426 assembling sequencing data into metagenomic contigs and grouping these contigs
427 into draft genomes (binning), yielding metagenome-assembled genomes (MAGs).
428 Notably, MAGs enable investigation of the genomes of uncultivatable organisms. While
429 MAGs can suffer from incompleteness and contamination due to limitations of
430 assembly and binning, software tools exist for evaluating MAG quality[63]. The majority of
431 publications to date have focused on creating MAGs from short-read sequencing
432 data[12,59,64], but generation of high-quality MAGs from long-read data from stool
433 samples has been recently reported[65]. To better characterize the genomes present in
434 our samples, we assembled and binned shotgun sequencing reads from South African
435 samples into MAGs. We generated 2419 MAGs (39 high-quality, 2038 medium-quality,
436 and 342 low-quality)[66] from 169 metagenomic samples (Supplementary Fig. 15A).
437 Applying the criteria for near-complete genomes proposed by Nayfach et al. (≥90%
438 complete, ≤5% contaminated, N50 ≥ 10 kb, average contig length ≥5 kb, ≤500
439 contigs, ≥90% of contigs with ≥5X read depth), 832 of these genomes (34%) are
440 designated near-complete. Filtering for completeness greater than 75% and
441 contamination less than 10% and de-replicating at 99% average nucleotide identity
442 (ANI) yielded a set of 1342 non-redundant medium-quality or better representative
443 strain genomes. This de-replicated collection includes VANISH taxa genomes,
444 including 94 *Prevotella*, 41 *Prevotellamassilia*, 39 *Succinivibrio*, and 10 Spirochaetota (4
445 *Treponema_D*, 6 *UBA9732*) (Fig. 5A, Supplementary Fig. 15, Supplementary Table 11).

446 To assess the novelty of this collection compared to known diversity of MAGs,
447 we compared our de-replicated MAG set to the Unified Human Gastrointestinal
448 Genome Collection (UHGG)[67]. Of these 1342 representative strain genomes, 16 (1.2%)
449 had less than 95% ANI to any genome in the full UHGG (Supplementary Fig. 15B) and
450 15 of these were retained in the final species set when de-replicated at 95% ANI
451 against UHGG species representatives (Supplementary Table 11) (two genomes with
452 less than 95% ANI to the UHGG species representatives were within 95% ANI of each
453 other and thus only one was retained after dereplication). These 15 genomes represent
454 7 GTDB phyla (Supplementary Fig. 15C) and 13 of 15 genomes (87%) are from
455 Bushbuckridge participants.
456 An additional 38 of 1332 genomes (2.9%) were not novel when compared to the
457 UHGG species representatives, but were assigned a higher genome quality score by
458 dRep than the corresponding UHGG representative (Supplementary Table 11, genome
459 scoring metrics in Methods and Olm et al. 2017). We note that ANI is calculated on the
460 basis of regions that align between genomes, and therefore may systematically
461 underestimate genomic novelty in this genome collection.
462 Interestingly, many MAGs within this set represent organisms that are
463 uncommon in Western microbiomes or not easily culturable, including organisms from
464 the genera *Treponema* and *Vibrio*. As short-read MAGs are typically fragmented and
465 exclude mobile genetic elements, we explored methods to create more contiguous
466 genomes, with a goal of trying to better understand these understudied taxa. We
467 performed long-read sequencing on three samples from participants in Bushbuckridge
468 with an Oxford Nanopore MinION sequencer (taxonomic composition of the three
469 samples shown in Supplementary Fig. 16). Samples were chosen for nanopore
470 sequencing on the basis of molecular weight distribution and total mass of DNA (see
471 Methods). One flow cell per sample generated an average of 19.71 Gbp of sequencing
472 with a read N50 of 8,275 bp after basecalling. From our three samples, we generated
473 741 nanopore MAGs (nMAGs), which yielded 35 non-redundant genomes when filtered
474 for completeness greater than 50% and contamination less than 10%, and de-
475 replicated at 99% ANI (Table 2, Supplementary Fig. 17, Supplementary Table 12).
476 Single-contig nMAGs were evaluated for GC skew to detect possible misassemblies.

477 All of the de-replicated nMAGs contained at least one full length 16S sequence, and
478 the contig N50 of 28 nMAGs was greater than 1 Mbp.
479 We compared assembly statistics between all MAGs and nMAGs, and found
480 that while nMAGs were typically less complete when evaluated by CheckM, the
481 contiguity of nanopore medium- and high-quality MAGs was an order of magnitude
482 higher (mean nMAG N50 of 260.5 kb compared to mean N50 of medium- and high-
483 quality MAGs of 15.1 kb) at comparable levels of average coverage (Supplementary
484 Fig. 17, Supplementary Fig. 18). We expect that CheckM under-calculates the
485 completeness of nanopore MAGs due to the homopolymer errors common in nanopore
486 sequencing, which result in frameshift errors when annotating genomes. Indeed, we
487 observe that nanopore MAGs with comparable high assembly size and low
488 contamination to short-read MAGs are evaluated by CheckM as having lower
489 completeness (Supplementary Fig. 18).
490
491 ***Novel genomes generated through nanopore sequencing***
492 When comparing the de-replicated medium- and high-quality nMAGs with the
493 corresponding short-read MAG for the same organism, we find that nMAGs typically
494 include many mobile genetic elements and associated genes that are absent from the
495 short-read MAG, such as transposases, recombinases, phages, and antibiotic
496 resistance genes (Figure 5A). Additionally, a number of the nMAGs are among the first
497 contiguous genomes in their clade. For example, we assembled two single contig,
498 megabase-scale genomes from the genus *Treponema,* a clade that contains various
499 commensal and pathogenic species. Notably, *Treponema* is a genus within the
500 Spirochaetes phylum, a VANISH taxa member that is often considered to be
501 completely lost with industrialization[9,11]. While some members of the genus are known
502 pathogens (*T. pallidum*), *Treponema* in non-industrialized communities is thought to
503 serve as a mutualistic fiber degrader in response to different fiber-rich nonwestern
504 diets[9]. The first of these genomes is a single-contig *Treponema succinifaciens* genome,
505 classified as *Treponema_D succinifaciens* by GTDB. The type strain of *T.*
506 *succinifaciens*, isolated from the swine gut[68], is the only genome of this species
507 currently available in public reference collections. Our *T. succinifaciens* genome is the

508 first complete genome of this species from the gut of a human. We assembled a
509 second *Treponema* sp. (GTDB *Treponema_D sp900541945*; Supplementary Fig. 19),
510 which contains a candidate natural product synthetic biosynthetic gene cluster (aryl
511 polyene cluster) and shares 92.1% ANI with *T. succinifaciens*. Additionally, we
512 assembled a 5.08 Mbp genome for Lentisphaerae *sp*., which has been shown to be
513 significantly enriched in traditional populations[69]. This genome also contains an aryl
514 polyene biosynthetic gene cluster and multiple beta-lactamases, and shares 94% 16S
515 rRNA identity with *Victivallis vadensis* and is classified as *Victivallis sp900550905* by
516 the GTDB, suggesting a new species or genus of the family *Victivallaceae* and
517 representing the second closed genome for the phylum Lentisphaerae.
518 Other nMAGs represent organisms that are prevalent in western individuals but
519 challenging to assemble due to their genome structure. Despite the prevalence of
520 *Bacteroides* in western microbiomes, only three closed *B. vulgatus* genomes are
521 available in RefSeq. We assembled a single contig, 2.68 Mbp *Bacteroides vulgatus*
522 (GTDB *Parabacteroides sp900549585*) genome that is 65.0% complete and 2.7%
523 contaminated and contains at least 16 putative insertion sequences, which may
524 contribute to the lack of contiguous short-read assemblies for this species. Similarly,
525 we assembled a single-contig genome for *Catabacter sp.*, a member of the order
526 *Clostridiales* (GTDB *CAG-475 sp900550915* of the Christensenellales order); the most
527 contiguous *Catabacter* genome in GenBank is in five scaffolded contigs[70]. The putative
528 *Catabacter sp.* shares 85% ANI with the best match in GenBank, suggesting that it
529 represents a new species within the *Catabacter* genus or a new genus entirely, and it
530 contains a sactipeptide biosynthetic gene cluster. Additionally, we assembled a 3.29
531 Mbp genome for *Prevotella sp.* (N50 = 1.14 Mbp), a highly variable genus that is
532 prevalent in nonwestern microbiomes and associated with a range of effects on host
533 health[71]. Notably, the first closed genomes of *P. copri*, a common species of *Prevotella*,
534 were only recently assembled with nanopore sequencing of metagenomic samples;
535 one from a human stool sample [65] and the other from cow rumen[72]. *P. copri* had
536 previously evaded closed assembly from short-read sequence data due to the dozens
537 of repetitive insertion sequences within its genome[65]. This *Prevotella* assembly contains

538 cephalosporin and beta-lactam resistance genes, as well as an aryl polyene
539 biosynthetic gene cluster.
540 Many long-read assembled genomes were evaluated to be of low completeness
541 despite having contig N50 values greater than 1 Mbp. Analysis showed that many of
542 these genomes had sparse or uneven short-read coverage, leading to gaps in short-
543 read polishing that would otherwise correct small frameshift errors. To polish genomic
544 regions that were not covered with short-reads, we performed long-read polishing on
545 assembled contigs from each sample, and re-binned polished contigs. Long-read
546 polishing improved the completeness of many organisms that are not commonly
547 described in the gut microbiota, due perhaps to their low relative abundance in the
548 average human gut, or to biases in shotgun sequencing library preparation that limit
549 their detection. For example, we generated a 2 Mbp Melainabacteria genome (GTDB
550 species *UMGS1477 sp900552205* of the family Gastranaerophilaceae).
551 Melainabacteria is a non-photosynthetic phylum closely related to Cyanobacteria that
552 has been previously described in the gut microbiome and is associated with
553 consuming a vegetarian diet[47]. Melainabacteria have proven difficult to isolate and
554 culture, and the only complete, single-scaffold genome existing in RefSeq was
555 assembled from shotgun sequencing of a human fecal sample[47]. Interestingly, our
556 Melainabacteria genome has a GC content of 30.9%, and along with assemblies of a
557 *Mycoplasma sp.* (GTDB *CAG_460 sp000437315* of class Bacilli) (25.3% GC) and
558 *Mollicutes sp.* (GTDB *Tener-01 sp001940985* of the class Bacilli) (28.1% GC)
559 (Supplementary Fig. 20), represent AT-rich organisms that can be underrepresented in
560 shotgun sequencing data due to the inherent GC bias of transposon insertion and
561 amplification-based sequencing approaches[73] (Supplementary Fig. 21, Supplementary
562 Fig. 22). Altogether, these three genomes increased in completeness by an average of
563 28.5% with long-read polishing to reach an overall average of 70.9% complete. While
564 these genomes meet the accepted standards to be considered medium-quality, it is
565 possible that some or all of these highly contiguous, megabase scale assemblies are
566 complete or near-complete yet underestimated by CheckM, for example due to
567 incomplete polishing.

568 Altogether, we find that *de novo* assembly approaches are capable of
569 generating contiguous, high-quality assemblies for novel organisms, offering potential
570 for investigation into the previously unclassified matter in the microbiomes of these
571 nonwestern communities. In particular, nanopore sequencing produced contiguous
572 genomes for organisms that are difficult to assemble due to repeat structures
573 (*Prevotella sp.*, *Bacteroides vulgatus*), as well as for organisms that are AT-rich
574 (*Mollicutes sp.*, *Melainabacteria sp.*). We observe that long-reads capture a broader
575 range of taxa both at the read and assembly levels when compared to short-read
576 assemblies, and that short- and long-read polishing approaches yield medium-quality
577 or greater draft genomes for these organisms. This illustrates the increased visibility
578 that *de novo* assembly approaches lend to the study of the full array of organisms in
579 the gut microbiome.
580

581 # Discussion

582 Together with Oduaran *et al.*, we provide the first description of gut microbiome
583 composition in Soweto and Bushbuckridge, South Africa, and to our knowledge, the
584 first effort utilizing shotgun and nanopore sequencing in South Africa to describe the
585 gut microbiome of adults. In doing so, we increase global representation in microbiome
586 research and provide a baseline for future studies of disease association with the
587 microbiome in South African populations, and in other transitional populations.

588 We find that gut microbiome composition differs demonstrably between the
589 Bushbuckridge and Soweto cohorts, further highlighting the importance of studying
590 diverse communities with differing lifestyle practices. Interestingly, even though gut
591 microbiomes of individuals in Bushbuckridge and Soweto share many features and are
592 more similar to each other than to other global cohorts studied, we do observe
593 hallmark taxa associated with westernization are enriched in microbiomes in Soweto.
594 These include *Bacteroides* and *Bifidobacterium*, which have been previously
595 associated with urban communities[3], consistent with Soweto's urban locale in the
596 Johannesburg metropolitan area.

597 We also observe enrichment in relative abundance of crAssphage and crAss-like
598 viruses in Soweto relative to Bushbuckridge, with relatively high prevalence in both
599 cohorts yet lower abundance on average of crAssphage clades alpha and delta
600 compared to several other populations. This furthers recent work which revealed that
601 crAssphage is prevalent across many cohorts globally[49], but found relatively fewer
602 crAssphage sequences on the African continent, presumably due to paucity of
603 available shotgun metagenomic data. Just as shotgun metagenomic sequence data
604 enables the study of viruses, it also enables us to assess the relative abundance of
605 human cells or damaged human cells in the stool. Surprisingly, we observe a high
606 relative abundance of human DNA in the raw sequencing data. We find a statistically
607 significantly higher relative abundance of human DNA in samples from Soweto
608 compared to those from Bushbuckridge. Future research may help illuminate the
609 potential reason for this finding, which may include a higher proportion of epithelium
610 disruption by invasive bacteria or parasites in Soweto vs. Bushbuckridge, and in South
611 Africa, in general, compared to other geographic settings. Alternatively, this may also

612 be attributable to a higher baseline of intestinal inflammation and fecal shedding of
613 leukocytes. Without additional information, it is difficult to speculate the reason for this
614 finding.

615 We find that individuals in Bushbuckridge are enriched in VANISH taxa including
616 *Succinatimonas*, which was recently reported to associate with microbiomes from
617 individuals practicing traditional lifestyles[12]. Intriguingly, several VANISH taxa
618 (*Succinatimonas*, *Succinivibrio*, *Treponema*) are bimodally distributed in the
619 Bushbuckridge cohort. We hypothesize that this bimodality could be caused by
620 differences in lifestyle and/or environmental factors including diet, history of
621 hospitalization or exposure to medicines, physical properties of the household
622 dwelling, differential treatment of drinking water across the villages comprising
623 Bushbuckridge. Additionally this pattern may be explained by participation in migration
624 to and from urban centers (or sharing a household with a migratory worker). A higher
625 proportion of men in the community engage in this pattern of rural-urban migration[39],
626 but it is possible that sharing a household with a cyclical worker could influence gut
627 microbiome composition via horizontal transmission[74].

628 Despite the fact that host genetics explain relatively little of the variation in
629 microbiome composition[75], we do observe a small number of taxa that associate with
630 host genetics in this population. Future work is required for replication and to
631 determine whether these organisms are interacting with the host and whether they are
632 associated with host health.

633 Additionally, we demonstrate marked differences between South African cohorts
634 and other previously studied populations living on the African continent and western
635 countries. Broadly, we find that South African microbiomes reflect the transitional
636 nature of their communities in that they overlap with western and nonwestern
637 populations. Tremendous human genetic diversity exists within Africa[76], and our work
638 reveals that there is a great deal of as yet unexplored microbiome diversity as well. In
639 fact, we find that microbiome beta diversity within communities may be systematically
640 underestimated by incomplete reference databases: taxa that are unique to individuals
641 in nonwestern populations are not present in reference databases and therefore not
642 included in beta diversity calculations. Though it has been reported that nonwestern

643 and traditional populations tend to have higher alpha diversity but lower beta diversity
644 compared to western populations, we show that this pattern is not universally upheld
645 when reference-agnostic nucleotide comparisons are performed. By extension, we
646 speculate that previous claims that beta diversity inversely correlates with alpha
647 diversity may have been fundamentally limited by study design in some cases.
648 Specifically, the disparity between comparing small, homogenous African populations
649 with large, heterogenous western ones constitutes a significant statistical confounder,
650 potentially preventing a valid assessment of beta diversity between groups.
651 Furthermore, alpha and beta diversity comparisons based on species-level taxonomic
652 assignment may be further confounded due to the presence of polyphyletic clades in
653 organisms like *Prevotella copri*[26,77] which are highly abundant in gut microbiomes of
654 nonwestern individuals. Notably, we also demonstrate that the notion of a "western-
655 nonwestern" axis of microbiome variation is over-simplified: we find taxa that are
656 enriched in South Africans relative to both western and hunter-gatherer/agriculturalist
657 cohorts.
658 Advances in sequencing technology are enhancing our ability to more
659 thoroughly characterize microbiomes using culture-free approaches. Through a
660 combination of short-read and long-read sequencing, we successfully assembled
661 contiguous, complete genomes for many organisms that are underrepresented in
662 reference databases, including genomes that are commonly considered to be enriched
663 in or limited to populations with traditional lifestyles including members of the VANISH
664 taxa (e.g., *Treponema sp., Treponema succinifaciens*). The phylum Spirochaetes,
665 namely its constituent genus *Treponema*, is considered to be a marker of traditional
666 microbiomes and has not been detected in high abundance in human microbiomes
667 outside of those communities[11,69]. Here, we identify Spirochaetes in the gut microbiome
668 of individuals in urban Soweto, demonstrating that this taxon is not exclusive to
669 traditional, rural populations, though we observe that relative abundance is higher on
670 average in traditional populations. Generation of additional genomes of VANISH taxa
671 and incorporation of these genomes into reference databases will allow for increased
672 sensitivity to detect these organisms in metagenomic data. Additionally, these
673 genomes facilitate comparative genomics of understudied gut microbes and allow for

674 functional annotation of potentially biologically relevant functional pathways. We note
675 that many of these genomes (e.g., Melainabacteria, *Succinatimonas*) are enriched in
676 the gut microbiota of Bushbuckridge participants relative to Soweto, highlighting the
677 impact of metagenomic assembly to better resolve genomes present in rural
678 populations.

679 In addition to investigating members of the VANISH taxa, long-read sequencing
680 enables the study of AT-rich genomes, which are difficult to sequence using
681 transposon-based library construction approaches common in short-read studies.
682 Thus, using long-read sequencing, we produced genomes for organisms that exist on
683 the extremes of the GC content spectrum, such as *Mycoplasma sp.*, *Mollicutes sp.*,
684 and *Melainabacteria sp*. We find that these organisms are sparsely covered by short-
685 read sequencing, illustrating the increased range of non-amplification based
686 sequencing approaches, such as nanopore sequencing. Interestingly, these
687 assemblies are evaluated as only medium-quality by CheckM despite having low
688 measurements of contamination, as well as genome lengths and gene counts
689 comparable to reference genomes from the same phylogenetic clade. We hypothesize
690 that sparse short-read coverage leads to incomplete polishing and therefore retention
691 of small frameshift errors, which are a known limitation of nanopore sequencing[78].
692 Further evaluation of 16S or long-read sequencing of traditional and western
693 populations can identify whether these organisms are specific to certain lifestyles, or
694 are more prevalent but poorly detected with shotgun sequencing.

695 While we find that the gut microbiome composition of the two South African
696 cohorts described herein reflects their lifestyle transition, we acknowledge that these
697 cohorts are not necessarily representative of all transitional communities in South
698 Africa or other parts of the world which differ in lifestyle, diet, and resource access.
699 Hence, further work remains to describe the gut microbiota in detail of other such
700 understudied populations. This includes a detailed characterization of parasites
701 present in microbiome sequence data, an analysis that we did not undertake in this
702 study but would be of great interest. These organisms have been detected in the
703 majority of household toilets in nearby KwaZulu-Natal province[79], and may interact with
704 and influence microbiota composition[80].

705 Our study has several limitations. Although the publicly available sequence data
706 from other global cohorts were generated with similar methodology to our study, it is
707 possible that batch effects exist between datasets generated in different laboratories
708 that may explain some percentage of the global variation we observe. Additionally,
709 while nanopore sequencing is able to broaden our range of investigation, we illustrate
710 that our ability to produce well-polished genomes at GC content extremes is limited.
711 This may affect our ability to accurately call gene lengths and structures, although
712 iterative long-read polishing improves our confidence in these assemblies. Future
713 investigation of these communities using less biased, higher coverage short-read
714 approaches or more accurate long-read sequencing approaches, such as PacBio
715 circular consensus sequencing, may improve assembly qualities. Additionally, long-
716 read sequencing of samples from a wider range of populations can identify whether the
717 genomes identified herein are limited to traditional and transitional populations, or
718 more widespread. Further, future improvements in error rate of long-read sequencing
719 may obviate the need for short-read polishing altogether.
720 Taken together, our results emphasize the importance of generating sequence
721 data from diverse transitional populations to contextualize studies of health and
722 disease in these individuals. To do so with maximum sensitivity and precision,
723 reference genomes must be generated to classify sequencing reads from these
724 metagenomes. Herein, we demonstrate the discrepancies in microbiome sequence
725 classifiability across global populations and highlight the need for more comprehensive
726 reference collections. Recent efforts have made tremendous progress in improving the
727 ability to classify microbiome data through creating new genomes via metagenomic
728 assembly[12,59,64], and here we demonstrate the application of short- and long-read
729 metagenomic assembly techniques to create additional genome references. Our
730 application of long-read sequencing technology to samples from South African
731 individuals has demonstrated the ability to generate highly contiguous MAGs and
732 shows immense potential to expand our reference collections and better describe
733 microbiomes throughout diverse populations globally. In the future, microbiome
734 studies may use a combination of short- and long-read sequencing to maximize
735 information output, perhaps performing targeted nanopore or other long-read

736 sequencing of samples that are likely to contain the most novelty on the basis of short-
737 read data.
738 The present study was conducted in close collaboration between site staff and
739 researchers in Bushbuckridge and Soweto as well as microbiome experts both in
740 South Africa and the United States, and community member feedback was invited and
741 incorporated at multiple phases in the planning and execution of the study (see
742 Oduaran *et al.* 2020 and Supplemental Information for additional detail). Tremendous
743 research efforts have produced detailed demographic and health characterization of
744 individuals living in both Bushbuckridge and Soweto[32,56,81,82] and it is our hope that
745 microbiome data can be incorporated into this knowledge framework in future studies
746 to uncover disease biomarkers or microbial associations with other health and lifestyle
747 outcomes. More broadly, we feel that this is an example of a framework for conducting
748 microbiome studies in an equitable manner, and we envision a system in which future
749 studies of microbiome composition can be carried out to achieve detailed
750 characterization of microbiomes globally while maximizing benefit to all participants
751 and researchers involved.
752

753 # Methods

754 ## Cohort selection

755 Stool samples were collected from women aged 40-72 years in Soweto, South
756 Africa and Bushbuckridge Municipality, South Africa. Participants were recruited on the
757 basis of participation in AWI-Gen[1], a previous study in which genotype and extensive
758 health and lifestyle survey data were collected. Human subjects research approval was
759 obtained (Stanford IRB 43069, University of the Witwatersrand Human Research Ethics
760 Committee M160121, Mpumalanga Provincial Health Research Committee
761 MP_2017RP22_851) and informed consent was obtained from participants for all
762 samples collected. Stool samples were collected and preserved in OmniGene Gut
763 OMR-200 collection kits (DNA Genotek). Samples were frozen within 60 days of
764 collection as per manufacturer's instructions, followed by long-term storage at -80°C.
765 As the enrollment criteria for our study included previous participation in a larger
766 human genomics project[1], we had access to self-reported ethnicity for each participant
767 (BaPedi, Ndebele, Sotho, Tsonga, Tswana, Venda, Xhosa, Zulu, Other, or Unknown).
768 Samples from participants who tested HIV-positive or who did not consent to an HIV
769 test were not analyzed.

770 ## Metagenomic sequencing of stool samples

771 DNA was extracted from stool samples using the QIAamp PowerFecal DNA Kit
772 (QIAGEN) according to the manufacturer’s instructions except for the lysis step, in
773 which samples were lysed using the TissueLyser LT (QIAGEN) (30 second
774 oscillations/3 minutes at 30Hz). DNA concentration of all DNA samples was measured
775 using Qubit Fluorometric Quantitation (DS DNA High-Sensitivity Kit, Life Technologies).
776 DNA sequencing libraries were prepared using the Nextera XT DNA Library Prep Kit
777 (Illumina). Final library concentration was measured using Qubit Fluorometric
778 Quantitation and library size distributions were analyzed with the Bioanalyzer 2100
779 (Agilent). Libraries were multiplexed and 150 base pair paired-end reads were
780 generated on the HiSeq 4000 platform (Illumina). Samples with greater than
781 approximately 300 ng remaining mass and a peak fragment length of greater than

782 19,000 bp (with minimal mass under 4,000 bp) as determined by a TapeStation 2200
783 (Agilent Technologies, Santa Clara, CA) were selected for nanopore sequencing.
784 Nanopore sequencing libraries were prepared using the 1D Genomic DNA by Ligation
785 protocol (ONT, Oxford UK) following standard instructions. Each library was sequenced
786 with a full FLO-MIN106D R9 Version Rev D flow cell on a MinION sequencer for at least
787 60 hours.

788 Literature review

789 Literature review criteria based on Brewster et al. 2019 [2] were employed: PubMed,
790 EMBASE, SCOPUS, and Web of Science were queried for observational and
791 interventional research involving the human gut microbiome through January 2021.
792 Terms including "gut microbiome" and "gut microbiota" and names of each of the 54
793 African countries were included in the search. Primary reports on the gut microbiome in
794 African children and/or adults, utilizing either 16S rRNA or shotgun metagenomic
795 sequencing and written in English, were included. Abstracts, secondary reports, poster
796 presentations, reviews or editorials, and *in vivo* and *in vitro* studies were excluded. The
797 list of relevant articles yielded by this search strategy was manually reviewed.

798 Computational methods

799 *Preprocessing and taxonomy profiling*
800 Stool metagenomic sequencing reads were trimmed using TrimGalore v0.6.5[3]
801 with a minimum quality score of 30 for trimming (--q 30) and minimum read length of
802 60 (--length 60). Trimmed reads were deduplicated to remove PCR and optical
803 duplicates using htstream SuperDeduper v1.2.0 with default parameters. Reads
804 aligning to the human genome (hg19) were removed using BWA v0.7.17-r1188[4].
805 Taxonomy profiles were created with Kraken v2.0.9-beta with default parameters[5] and
806 (i) a comprehensive custom reference database containing all bacterial and archaeal
807 genomes in GenBank assembled to "complete genome," "chromosome," or "scaffold"
808 quality as of January 2020, and (ii) the pre-built Struo[6] GTDB release 95 database
809 containing one genome per species. Bracken v2.2.0 was then used to re-estimate

810 abundance at each taxonomic rank[7]. MetaPhlAn3[8] taxonomy profiles were also
811 generated.
812
813 *Additional data*
814 Published data from additional adult populations were downloaded from the
815 NCBI Sequence Read Archive (SRA) or European Nucleotide Archive (Supplementary
816 Table 9) and preprocessed and taxonomically classified as described above. The study
817 by Backhed et al. sampled both mothers and infants: only the maternal samples were
818 retained in this study. For datasets containing longitudinal samples from the same
819 individual, one unique sample per individual was chosen (the first sample from each
820 individual was chosen from the United States Human Microbiome Project cohort).
821
822 *K-mer sketches*
823 *K*-mer sketches were computed using sourmash v2.0.0[9]. Low abundance *k*-
824 mers were trimmed using the "trim-low-abund.py" script from the khmer package[10]
825 with a *k*-mer abundance cutoff of 3 (-C 3) and trimming coverage of 18 (-Z 18).
826 Signatures were computed for each sample using the command "sourmash compute"
827 with a compression ratio of 1000 (--scaled 1000) and *k*-mer lengths of 21, 31, and 51 (-
828 k 21,31,51). Two signatures were computed for each sample - one signature tracking
829 *k*-mer abundance (--track-abundance flag) for angular distance comparisons, and one
830 without this flag for Jaccard distance comparisons. Signatures at each length of *k* were
831 compared using "sourmash compare" with default parameters and the correct length
832 of *k* specified with the -k flag.
833
834 *Functional annotation*
835 Unassembled metagenomic reads were functionally profiled using ShortBRED[11]
836 v0.9.3 with a pre-built antibiotic resistance database based on the Comprehensive
837 Antibiotic Resistance Database[12]. Features were pre-filtered for >10% prevalence and
838 statistical analysis was performed using MaAsLin v2[13] using the compound Poisson
839 linear model (CPLM) and total sum scaling normalization with "site" as a fixed effect.

840 Pangenomes were calculated with PanPhlAn v3.1 [8] using parameters for
841 increased sensitivity recommended by the authors of the tool: "--min_coverage 1 --
842 left_max 1.70 --right_min 0.30".
843 MetaCyc pathways were profiled with HUMAnN v3.0.0[8] with default parameters,
844 using the mpa_v30_CHOCOPhlAn_201901 database. Forward and reverse reads were
845 concatenated into one file per sample prior to processing. Pathway abundances were
846 normalized to copies per million (CPM) and statistical analysis was performed using
847 MaAsLin v2 using the compound Poisson linear model (CPLM) and total sum scaling
848 normalization with "site" as a fixed effect.
849
850 *Genome assembly, binning, and evaluation*
851 Short-read metagenomic data were assembled with SPAdes v3.15[14] and binned
852 into draft genomes using a publicly available workflow
853 (https://github.com/bhattlab/bhattlab_workflows/blob/master/binning/bin_das_tool_ma
854 nysamp.snakefile, commit version bbe6511 as of Apr 20, 2021). Briefly, short reads
855 were aligned to assembled contigs with BWA v0.7.17[4] and contigs were subsequently
856 binned into draft genomes with MetaBAT v2.15[15], CONCOCT v1.1.0 [16], and MaxBin
857 v2.2.7[17]. Default parameters were used for each binner, with the following exceptions:
858 For the jgi_summarize_bam_contig_depths step of MetaBAT, minimum contig length
859 was set at 1000 bp (--minContigLength 1000), minimum contig depth of coverage of 1
860 (--minContigDepth 1), and a minimum end-to-end percent identity of reads of 50 (--
861 percentIdentity 50). Bins were aggregated and refined with DASTool v1.1.1[18]. Bins were
862 evaluated for size, contiguity, completeness, and contamination with QUAST v5.0.2[19],
863 CheckM v1.0.13[20], Prokka v1.14.6[21], Aragorn v1.2.38[22], and Barrnap v0.9
864 (https://github.com/tseemann/barrnap/). We referred to published guidelines to
865 designate genome quality[23]. Individual contigs from all assemblies were assigned
866 taxonomic classifications with Kraken v2.0.9[5,23]. To create de-replicated genome
867 collections, genomes with completeness greater than 75% and contamination less
868 than 10% (as evaluated by CheckM) were de-replicated using dRep v3.2.0[24] with ANI
869 threshold to form secondary clusters (-sa) at 0.99 (strain-level) or 0.95 (species-level).
870 For comparison to UHGG species representatives secondary ANI was set to 0.95.

871 dRep chooses the genome with the highest score as the cluster representative
872 according to the following formula: *dRep score = A*Completeness - B*Contamination +*
873 *C*(Contamination*(Strain heterogeneity / 100)) + D*log(N50) + E*log(size) + F*(centrality-*
874 *secondary ani)*. A through F are values which can be tuned by the user to change the
875 relative importance of each parameter in choosing representative genomes. Default
876 parameters (A=1, B=5, C=1, D=0.5, E=0, F=1) were used herein.
877 Long-read data were assembled with Lathe[25] as previously described. Briefly,
878 Lathe implements basecalling with Guppy v2.3.5, assembly with Flye v2.4.2[26], short-
879 read polishing with Pilon v1.23[27], and circularization with Circlator[28] and Encircle[25].
880 Contigs greater than 1,000 bp were subsequently binned into draft genomes with
881 MetaBAT v2.13 using minimum contig depth coverage of 1, minimum end-to-end
882 percent identity of reads of 50, and otherwise using default parameters, then classified,
883 and de-replicated as described above. Additional long-read polishing was performed
884 using four iterations of polishing with Racon v1.4.10[29] and long-read alignment using
885 minimap2 v2.17-r941[30], followed by one round of polishing with Medaka v0.11.5
886 (https://github.com/nanoporetech/medaka). Single-contig genomes were analyzed for
887 GC skew using SkewIT[31]. Genomes of interest were plotted with the DNAPlotter GUI [32].
888 Draft genomes were additionally classified with GTDBtk v1.4.1 (classify_wf)[33]
889 using release 95 reference data.
890 Direct comparisons between nMAGs and corresponding MAGs were performed
891 by de-replicating high- and medium-quality nMAGs with MAGs assembled from the
892 same sample. MAGs sharing at least 99% ANI with an nMAG were aligned to the
893 nMAG regions using nucmer v3.1 and uncovered regions of the nMAG were annotated
894 with prokka 1.14.6, VIBRANT v1.2.1[34], and ResFams v1.2[35].
895 Phylogenetic trees for all dereplicated short- and long-read MAGs were
896 constructed with GTDBtk v1.4.1. To construct phylogenetic trees for taxa of interest,
897 reference 16S sequences were downloaded from the Ribosomal Database Project
898 (Release 11, update 5, September 30, 2016)[37] and 16S sequences were identified from
899 nanopore genome assemblies using Barrnap v0.9
900 (https://github.com/tseemann/barrnap/). Sequences were aligned with MUSCLE
901 v3.8.1551[38] with default parameters. Maximum-likelihood phylogenetic trees were

902 constructed from the alignments with FastTree v2.1.10[38,39] with default settings (Jukes-
903 Cantor + CAT model). Support values for branch splits were calculated using the
904 Shimodaira-Hasegawa test with 1,000 resamples (default). Trees were visualized with
905 FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).
906
907 *Statistical analysis and plotting*
908 Statistical analyses were performed using R v4.0.2[40] with packages MASS v7.3-53[41],
909 stats[40], ggsignif v0.6.0[42], and ggpubr v0.4.0[43]. Alpha and beta diversity were calculated
910 using the vegan package v2.6.0[44]. Two-sided Wilcoxon rank-sum tests were used to
911 compare alpha and beta diversity between cohorts. Count data were rarefied and
912 normalized via cumulative sum scaling and log2 transformation[45] prior to MDS. Data
913 separation in MDS was assessed via PERMANOVA (permutation test with pseudo F
914 ratios) using the adonis function from the vegan package. Differential microbial features
915 between individuals living in Soweto and Bushbuckridge were identified from
916 unnormalized count data output from kraken2 classification and bracken abundance
917 re-estimation and filtered for 20% prevalence and at least 500 sequencing reads using
918 DESeq2 with the formula "~site"[46]. Plots were generated in R using the following
919 packages: cowplot v1.0.0[47], DESeq2 v1.28.0[46], genefilter v1.70.0[48], ggplot2 v3.3.2[49],
920 ggpubr v0.4.0, ggrepel v0.8.2[50], ggsignif v0.6.0, gtools v3.8.2[51], harrietr v0.2.3[52], MASS
921 v7.3-53, reshape2 v1.4.4[53], tidyverse v1.3.0[54], and vegan v2.6.0.

922 Data availability

923 All shotgun sequence data generated by this study, as well as metagenome-
924 assembled genome sequences are deposited in the NCBI Sequence Read Archive
925 under BioProject PRJNA678454. Participant-level metadata (age, BMI, blood pressure
926 measurements, and concomitant medications) and human genetic data will be
927 deposited in the European Genome-phenome Archive (EGA) under Study ID
928 EGAS00001002482 and dataset ID EGAD0000100658.

Code availability

R code for analysis and figure generation is available at https://github.com/bhattlab/SouthAfrica. Data analysis workflows referenced in Methods are available at https://github.com/bhattlab/bhattlab_workflows.

Acknowledgements

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

This work was supported in part by a grant from the Stanford Center for Innovation in Global Health and by NIH grant P30 CA124435, which supports the following Stanford Cancer Institute Shared Resource: the Genetics Bioinformatics Service Center. A.S.B was supported by the Rosenkranz prize and by an R01AI148623 from the National Institute of Allergy and Infectious Diseases. F.B.T. was supported by the National Science Foundation Graduate Research Fellowship and the Stanford Computational, Evolutionary, and Human Genetics Pre-Doctoral Fellowship. D.G.M was supported by the Stanford Graduate Fellowships in Science and Engineering program. O.H.O. was partially supported by a Fogarty Global Health Equity Scholar award (TW009338). ANW is supported by the Fogarty International Centre, National Institutes of Health under award number K43TW010698. The work was further

supported by the South African National Research Foundation (CPRR160421162721) and a seed grant from the African Partnership for Disease Control. The AWI-Gen project is supported by the National Human Genome Research Institute (U54HG006938) as part of the H3A Consortium. The MRC/Wits Rural Public Health and Health Transitions Research Unit and Agincourt Health and Socio-Demographic Surveillance System, a node of the South African Population Research Infrastructure Network (SAPRIN), is supported by the Department of Science and Innovation, the University of the Witwatersrand and the Medical Research Council, South Africa, and previously the Wellcome Trust, UK (grants 058893/Z/99/A; 069683/Z/02/Z; 085477/Z/08/Z; 085477/B/08/Z).This paper describes the views of the authors and does not necessarily represent the official views of the National Institutes of Health (USA).

Author contributions

A.S.B. and S.H. conceived of study and secured funding. A.S.B., S.H., Z.N., R.T., X.G.O., F.W., A.N.W., R.G.W., K.K., S.T., S.A.N., and V.S. organized study logistics and coordinated participant enrollment and sample collection. V.S., M.R.H., O.H.O., F.B.T., and R.B. contributed to sample preparation and sequencing. F.B.T., D.M., and S.H. performed data analysis. A.S.B., F.B.T., and D.M. wrote and edited the manuscript, S.H. edited the manuscript.

Competing interests

The authors declare no competing interests.

Main Tables

Table 1. Participant characteristics

	Site	Mean	Standard deviation	Range
Age	**Bushbuckridge**	55.2	7.9	43.0 - 72.0
	Soweto	54.1	5.9	43.0 - 64.0
BMI*	**Bushbuckridge**	32.4	8.0	21.2 - 59.0
	Soweto	36.1	9.3	20.4 - 58.6
Systolic blood pressure**	**Bushbuckridge**	137.0	18.3	101.3 - 189.3
	Soweto	134.5	22.5	96.0 - 193.0
Diastolic blood pressure**	**Bushbuckridge**	83.7	12.1	54.0 - 119.0
	Soweto	90.0	14.4	58.0 - 119.0

*One Bushbuckridge participant's BMI measurement was excluded as the recorded value was 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.
**A second participant from Bushbuckridge had missing blood pressure measurements and is not included in blood pressure summary statistics

990

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

992

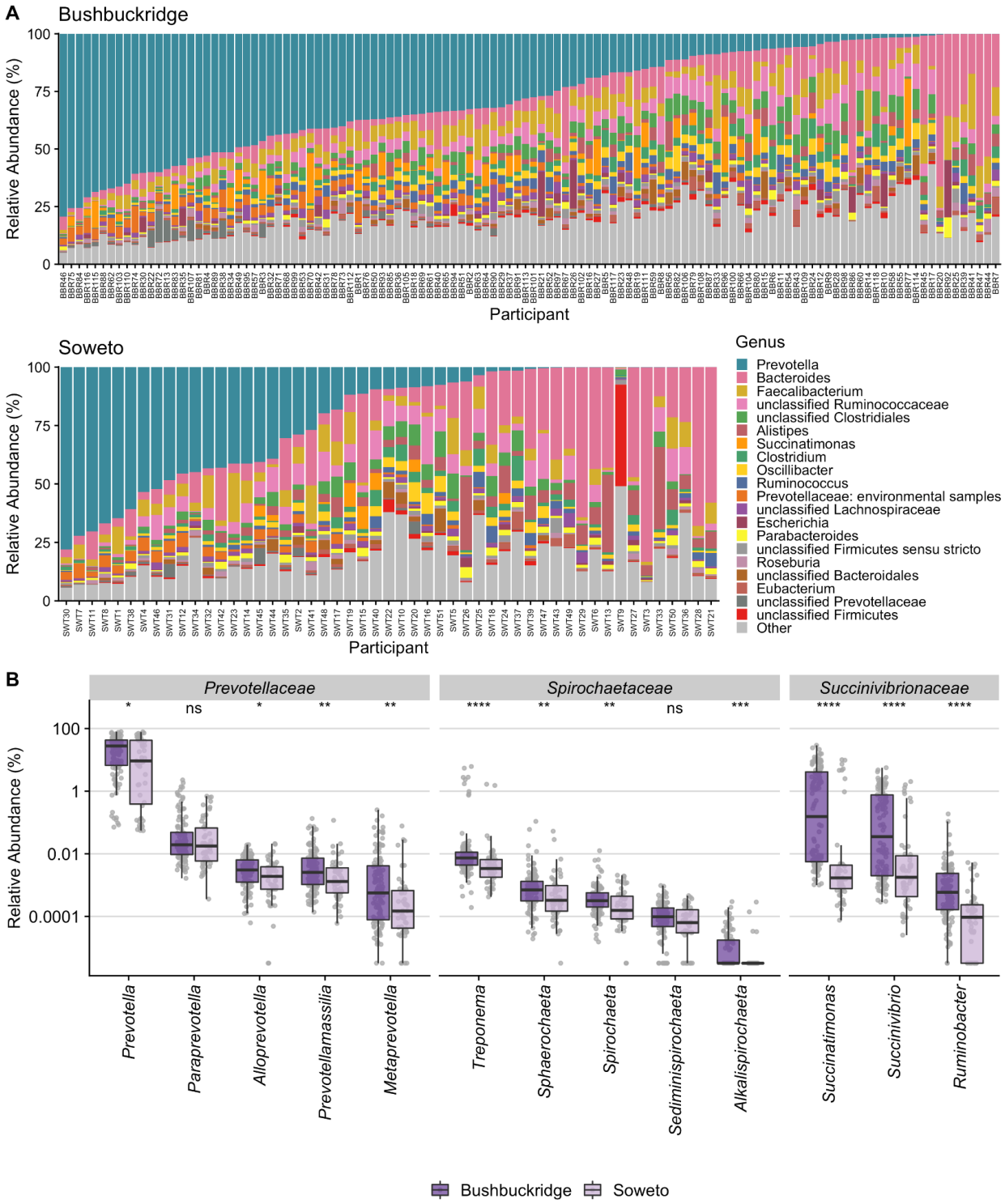
Classification	GTDB Classification	Size (Mb)	Contigs	N50 (Mb)	Quality	16S rRNAs	GC %	GC Skew	Polishing
Alistipes putredinis	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Rikenellaceae;g__;s__	1.91	1	1.91	Medium-quality	2	53.1	0.96	Short Read Only
Anaerotruncus sp.	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Acutalibacteraceae;g__Eubacterium_R;s__Eubacterium_R sp000433975	2.04	1	2.04	Medium-quality	2	43.71	0.94	Short Read Only
Bacilli bacterium	d__Bacteria;p__Firmicutes;c__Bacilli;o__RF39;f__CAG-302;g__CAG-302;s__CAG-302 sp900548425	1.46	1	1.46	Medium-quality	1	26.19	0.93	Short Read Only
Bacteroidales bacterium	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Paludibacteraceae;g__RF16;s__RF16 sp900556095	2.67	2	1.8	High-quality	3	47.31	NA	Short Read Only
Bacteroidales bacterium	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Muribaculaceae;g__CAG-279;s__CAG-279 sp000437795	2.79	1	2.79	High-quality	4	49.82	0.92	Short Read Only
Bacteroidales bacterium	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Rikenellaceae;g__Alistipes;s__Alistipes sp900546065	1.7	1	1.7	Medium-quality	1	56.6	0.7	Short Read Only
Bacteroides sp.	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__UBA932;g__RC9;s__RC9 sp000432655	2	2	1.59	High-quality	3	48.24	NA	Short Read Only
Bacteroides sp.	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Phocaeicola;s__Phocaeicola sp000434735	2.82	2	2	Medium-quality	6	43.31	NA	Short Read Only
Bacteroides vulgatus	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Tannerellaceae;g__Parabacteroides;s__Parabacteroides sp900549585	2.68	1	2.68	Medium-quality	3	42.71	0.84	Short Read Only
Candidatus Melainabacteria	d__Bacteria;p__Cyanobacteria;c__Vampirovibrionia;o__Gastranaerophilales;f__Gastranaerophilaceae;g__UMGS1477;s__UMGS1477 sp900552205	2	1	2	Medium-quality	1	30.9	0.32	Short and Long
Catabacter sp.	d__Bacteria;p__Firmicutes_A;c__Clostridia_A;o__Christensenellales;f__CAG-917;g__CAG-475;s__CAG-475 sp900550915	1.65	1	1.65	Medium-quality	1	46.4	0.87	Short and Long
Clostridiales bacterium	d__Bacteria;p__Firmicutes_A;c__Clostridia_A;o__Christensenellales;f__CAG-74;g__UBA11524;s__UBA11524 sp000437595	2.03	4	0.6	Medium-quality	4	57.9	NA	Short Read Only
Clostridiales bacterium	d__Bacteria;p__Firmicutes_A;c__Clostridia_A;o__Christensenellales;f__CAG-917;g__CAG-349;s__CAG-349 sp003539515	1.53	1	1.53	Medium-quality	1	47.28	0.94	Short Read Only
Clostridiales bacterium	d__Bacteria;p__Firmicutes_A;c__Clostridia_A;o__Christensenellales;f__CAG-138;g__PeH17;s__PeH17 sp000435055	1.95	4	0.73	Medium-quality	3	49.59	NA	Short Read Only
Clostridiales bacterium	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__CAG-272;g__CAG-724;s__	2.24	5	0.58	Medium-quality	2	48.65	NA	Short Read Only
Clostridiales bacterium	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__UMGS1810;f__UMGS1810;g__;s__	2.65	1	2.65	Medium-quality	3	42.82	0.69	Short Read Only

Clostridiales bacterium	d__Bacteria;p__Firmicutes_A;c__Clostridia_A;o__Christensenellales;f__Borkfalkiaceae;g__UBA1259;s__	1.32	2	0.79	Medium-quality	1	45.19	NA	Short Read Only
Clostridiales bacterium	d__Bacteria;p__Firmicutes_A;c__Clostridia_A;o__Christensenellales;f__CAG-917;g__CAG-349;s__CAG-349 sp003539515	1.61	1	1.61	Medium-quality	1	46.9	0.94	Short Read Only
Clostridium sp.	d__Bacteria;p__Firmicutes;c__Bacilli;o__RF39;f__CAG-1000;g__CAG-1000;s__	1.53	1	1.53	Medium-quality	1	25.24	0.89	Short Read Only
Clostridium sp.	d__Bacteria;p__Firmicutes_A;c__Clostridia_A;o__Christensenellales;f__CAG-917;g__CAG-349;s__CAG-349 sp003539515	1.3	1	1.3	Medium-quality	1	46.87	0.8	Short Read Only
Clostridium sp.	d__Bacteria;p__Firmicutes;c__Bacilli;o__Acholeplasmatales;f__Anaeroplasmataceae;g__;s__	2.01	1	2.01	Medium-quality	3	28.81	0.92	Short Read Only
Clostridium sp.	d__Bacteria;p__Firmicutes;c__Bacilli;o__RF39;f__CAG-1000;g__CAG-533;s__CAG-533 sp000434495	1.14	1	1.14	Medium-quality	1	29.09	0.7	Short Read Only
Clostridium sp.	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Acutalibacteraceae;g__CAG-177;s__CAG-177 sp000431775	2.44	2	2.23	High-quality	3	52.53	NA	Short Read Only
Eubacterium	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Acutalibacteraceae;g__UMGS1532;s__UMGS1532 sp900552605	2	4	0.63	Medium-quality	2	44.52	NA	Short Read Only
Lachnospiraceae bacterium	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__CAG-95;s__	3.38	2	1.94	Medium-quality	4	43.55	NA	Short Read Only
Lachnospiraceae bacterium	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__CAG-95;s__CAG-95 sp000438155	3.81	7	2.83	Medium-quality	4	43.56	NA	Short Read Only
Lentisphaeria bacterium	d__Bacteria;p__Verrucomicrobiota;c__Lentisphaeria;o__Victivallales;f__Victivallaceae;g__Victivallis;s__Victivallis sp900550905	5.08	1	5.08	Medium-quality	3	57.5	0.69	Short and Long
Mollicutes bacterium	d__Bacteria;p__Firmicutes;c__Bacilli;o__ML615J-28;f__CAG-698;g__Tener-01;s__Tener-01 sp001940985	1.68	2	1.49	Medium-quality	2	28.1	NA	Short and Long
Mycoplasma sp.	d__Bacteria;p__Firmicutes;c__Bacilli;o__RF39;f__CAG-1000;g__CAG-460;s__CAG-460 sp000437315	1.17	3	1.12	Medium-quality	2	25.3	NA	Short and Long
Oscillibacter sp.	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Oscillospiraceae;g__Oscillibacter;s__Oscillibacter sp001916835	1.13	10	0.17	Medium-quality	1	57.37	NA	Short Read Only
Porphyromonadaceae bacterium	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Muribaculaceae;g__C941;s__C941 sp004557565	2.97	1	2.97	Medium-quality	5	47.43	0.76	Short Read Only
Prevotella sp.	d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Prevotella;s__Prevotella sp000434515	3.29	3	1.14	Medium-quality	6	43.6	NA	Short and Long
Ruminococcaceae bacterium	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Acutalibacteraceae;g__Ruminococcus_E;s__Ruminococcus_E sp003526955	1.95	3	0.8	Medium-quality	4	38.35	NA	Short Read Only
Ruminococcaceae bacterium	d__Bacteria;p__Firmicutes_A;c__Clostridia_A;o__Christensenellales;f__QALW01;g__UMGS1338;s__UMGS1338 sp900550805	2.27	1	2.27	High-quality	3	51.43	0.91	Short Read Only
Ruminococcaceae bacterium	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__CAG-272;g__CAG-448;s__	1.78	1	1.78	Medium-quality	3	58.25	0.63	Short Read Only
Treponema sp.	d__Bacteria;p__Spirochaetota;c__Spirochaetia;o__Treponematales;f__Trepo	2.06	1	2.06	Medium-quality	3	41.55	0.93	Short Read Only

	nemataceae;g__Treponema_D;s__Treponema_D sp900541945								
Treponema succinifaciens	d__Bacteria;p__Spirochaetota;c__Spirochaetia;o__Treponematales;f__Treponemataceae;g__Treponema_D;s__Treponema_D succinifaciens	2.55	1	2.55	High-quality	4	39.12	0.82	Short Read Only
uncultured Ruminococcus	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Acutalibacteraceae;g__CAG-180;s__CAG-180 sp004556705	1.59	2	1.34	Medium-quality	2	44	NA	Short Read Only
uncultured Ruminococcus	d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Oscillospirales;f__Ruminococcaceae;g__CAG-353;s__CAG-353 sp900066885	2.08	1	2.08	Medium-quality	5	46.85	0.69	Short Read Only

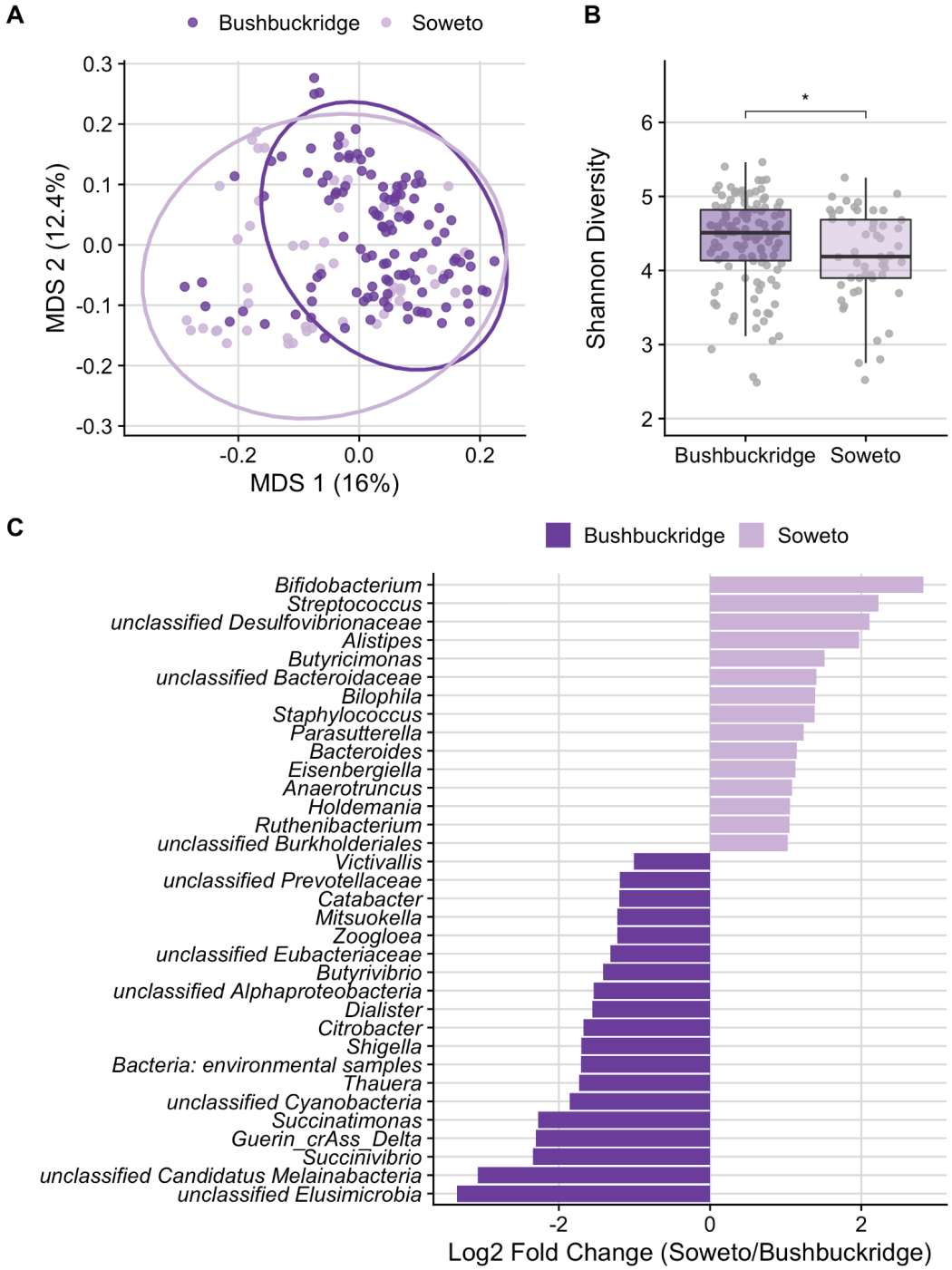
993

Figures


A
Bushbuckridge
Relative Abundance (%)
Participant
Soweto
Relative Abundance (%)
Participant
Genus
Prevotella
Bacteroides
Faecalibacterium
unclassified Ruminococcaceae
unclassified Clostridiales
Alistipes
Succinatimonas
Clostridium
Oscillibacter
Ruminococcus
Prevotellaceae: environmental samples
unclassified Lachnospiraceae
Escherichia
Parabacteroides
unclassified Firmicutes sensu stricto
Roseburia
unclassified Bacteroidales
Eubacterium
unclassified Prevotellaceae
unclassified Firmicutes
Other
B
Prevotellaceae
Spirochaetaceae
Succinivibrionaceae
Relative Abundance (%)
Prevotella
Paraprevotella
Alloprevotella
Prevotellamassilia
Metaprevotella
Treponema
Sphaerochaeta
Spirochaeta
Sediminispirochaeta
Alkalispirochaeta
Succinatimonas
Succinivibrio
Ruminobacter
Bushbuckridge
Soweto

997 **Figure 1. Taxonomic composition of South African study participants**
998 Sequence data were taxonomically classified using Kraken2 with a database
999 containing all genomes in GenBank of "scaffold" quality or better as of January 2020.
1000 (A) Top 20 genera by relative abundance for samples from participants in
1001 Bushbuckridge and Soweto, sorted by decreasing *Prevotella* abundance. *Prevotella*,
1002 *Faecalibacterium*, and *Bacteroides* are the most prevalent genera across both study
1003 sites.
1004 (B) Relative abundance of VANISH genera by study site, grouped by family. A
1005 pseudocount of 1 read was added to each sample prior to relative abundance
1006 normalization in order to plot on a log scale, as the abundance of some genera in some
1007 samples is zero. Relative abundance values of most VANISH genera are higher on
1008 average in participants from Bushbuckridge than Soweto (Two-sided Wilcoxon rank-
1009 sum test, significance values denoted as follows: (*) $p < 0.05$, (**) $p < 0.01$, (***) $p <$
1010 0.001, (****) $p < 0.0001$, (ns) not significant). For box plots, lower and upper hinges
1011 correspond to the first and third quartiles, upper and lower box plot whiskers represent
1012 the highest and lowest values within 1.5 times the interquartile range, and the
1013 horizontal line represents the median.

1014



1015

1016 **Figure 2. Comparison of Bushbuckridge and Soweto microbiomes**

1017 (A) Multidimensional scaling of pairwise Bray-Curtis distance between samples

1018 (rarefied to 1.44M counts per sample to control for read depth and CSS normalized).

1019 Soweto samples have greater dispersion than Bushbuckridge (PERMDISP2 $p < 0.001$).

1020 (B) Shannon diversity calculated on rarefied species-level taxonomic classifications for
1021 each sample. Samples from Bushbuckridge are higher in alpha diversity than samples
1022 from Soweto (Two-sided Wilcoxon rank-sum test, $p < 0.05$). For box plots, lower and
1023 upper hinges correspond to the first and third quartiles, upper and lower box plot
1024 whiskers represent the highest and lowest values within 1.5 times the interquartile
1025 range, and the horizontal line represents the median.
1026 (C) DESeq2 identifies microbial genera that are differentially abundant in rural
1027 Bushbuckridge compared to the urban Soweto cohort. Features with log2 fold change
1028 greater than one are plotted (full results in Supplementary Table 7).
1029

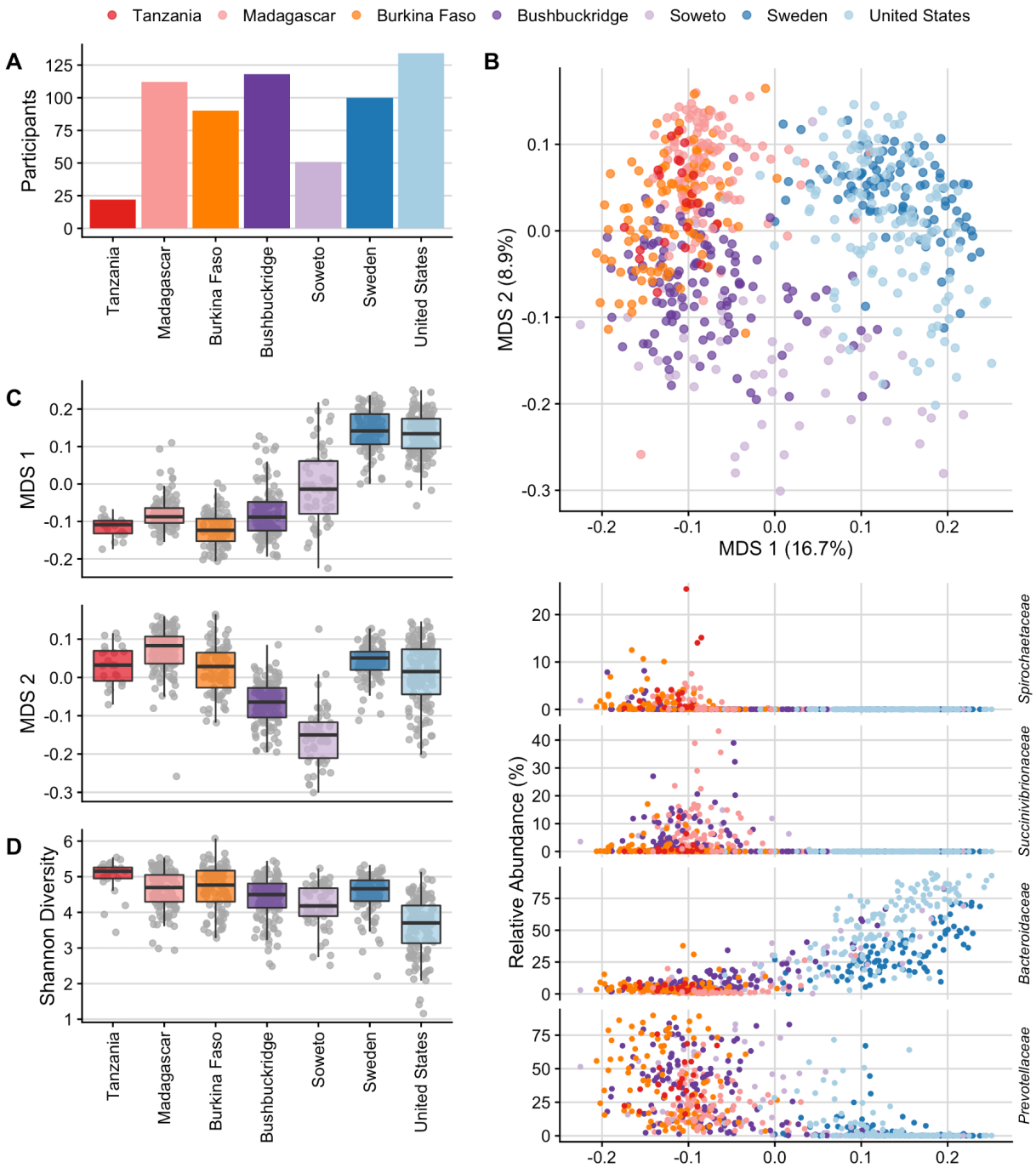


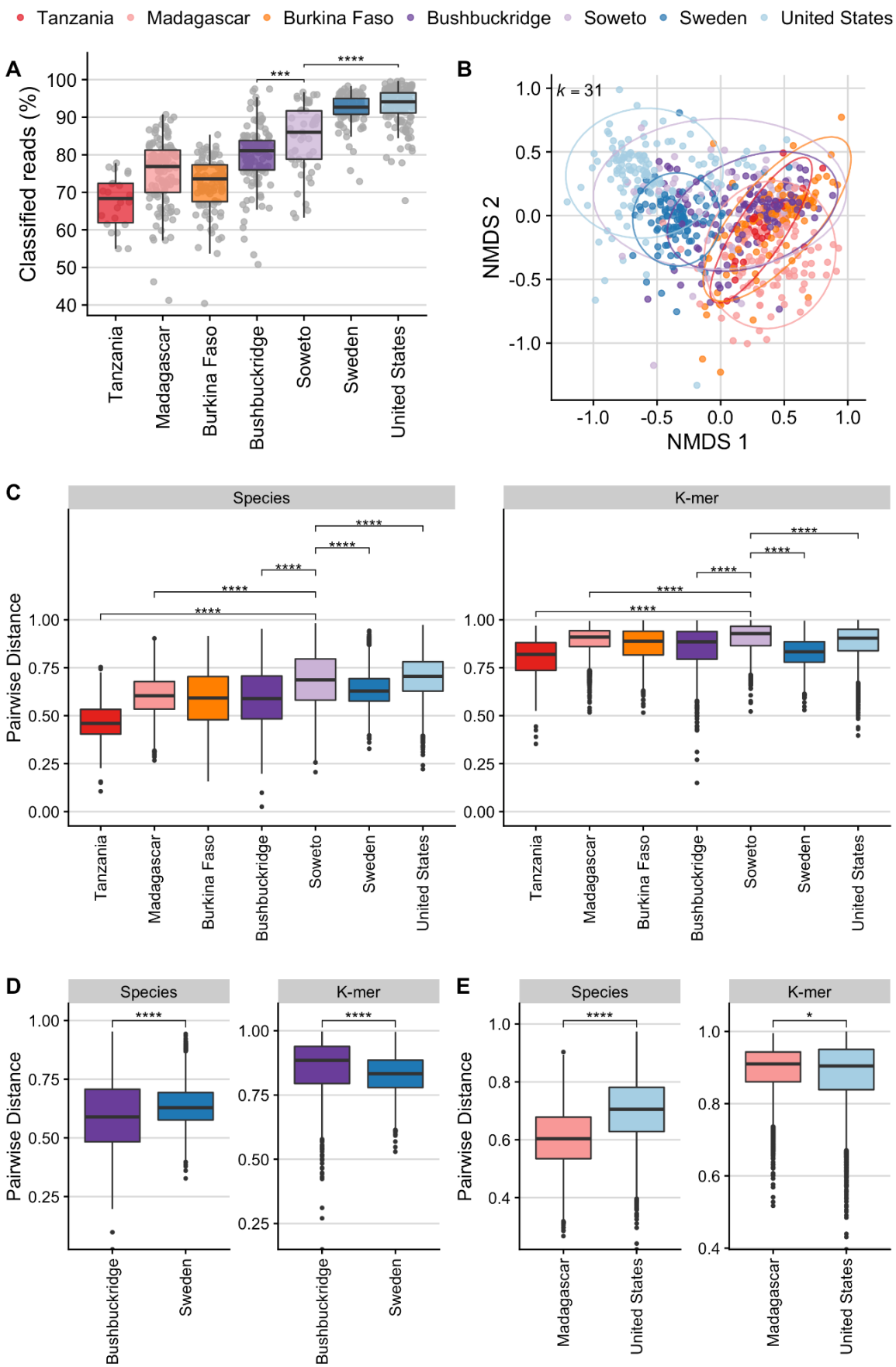
Figure 3. Community-level comparison of global microbiomes

Comparisons of South African microbiome data to microbiome sequence data from four publicly available cohorts representing western (United States, Sweden) and nonwestern (Tanzania, Madagascar, Burkina Faso) 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

1038 populations (Sweden, United States) cluster away from African populations practicing a
1039 traditional lifestyle (Madagascar, Tanzania, Burkina Faso) while transitional South
1040 African microbiomes overlap with both western and nonwestern populations. Shown
1041 below are scatterplots of relative abundance of the top four taxa most correlated with
1042 MDS 1 (Spearman's rho, *Spirochaetaceae* -0.824, *Succinivibrionaceae* -0.804,
1043 *Bacteroidaceae* 0.769, and *Prevotellaceae* -0.752) against MDS 1 on the x-axis.
1044 (C) Box plots of the first axis of MDS (MDS 1) which correlates with geography and
1045 lifestyle, and the second axis of MDS (MDS 2) which shows a distinct separation of
1046 South African cohorts.
1047 (D) Shannon diversity across cohorts. Shannon diversity was calculated from data
1048 rarefied to the number of counts of the lowest sample.
1049 For box plots in (C) and (D), lower and upper hinges correspond to the first and third
1050 quartiles, upper and lower box plot whiskers represent the highest and lowest values
1051 within 1.5 times the interquartile range, and the horizontal line represents the median.
1052
1053


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

1055 **Figure 4. Comparison of beta diversity between communities calculated by**
1056 **taxonomy versus nucleotide *k*-mer composition**
1057 (A) Percentage of reads classified at any taxonomic rank, by cohort, based on a
1058 reference database of all reference genomes of "scaffold" quality or higher in GenBank
1059 and RefSeq as of January 2020. Western microbiomes have a higher percentage of
1060 classifiable reads compared to nonwestern microbiomes (Two-sided Wilcoxon rank-
1061 sum test $p < 0.001$).
1062 (B) Nucleotide sequences of microbiome sequencing reads were compared using *k*-
1063 mer sketches. This reference-free approach is not constrained by comparison to
1064 existing genomes and therefore allows direct comparison of sequences. Briefly, a hash
1065 function generates signatures at varying sequence lengths (*k*) and *k*-mer sketches can
1066 be compared between samples. Data shown here are generated from comparisons at
1067 *k*=31 (approx. species-level)[61]. Non-metric multidimensional scaling (NMDS) of angular
1068 distance values computed between each pair of samples.
1069 (C-E) Comparison of pairwise beta diversity within communities assessed by Bray-
1070 Curtis distance based on species-level classifications and angular distance of
1071 nucleotide *k*-mer sketches. (C) All populations. (D) South African populations
1072 (Bushbuckridge and Soweto) compared to the Swedish cohort. Beta diversity
1073 measured by Bray-Curtis distance is higher in Soweto but lower in Bushbuckridge
1074 compared to the United States. However, reference-independent *k*-mer comparisons
1075 indicate that nucleotide dissimilarity is higher within both South African populations
1076 compared to the Swedish cohort. (E) Species-based Bray-Curtis distance indicates
1077 that there is more beta diversity within the United States cohort compared to
1078 Malagasy, but *k*-mer distance indicates an opposite pattern.
1079 For all box plots in (A), (C), (D), and (E), lower and upper hinges correspond to the first
1080 and third quartiles, upper and lower box plot whiskers represent the highest and lowest
1081 values within 1.5 times the interquartile range, and the horizontal line represents the
1082 median. Significance values for two-sided Wilcoxon rank sum tests denoted as follows:
1083 (*) $p < 0.05$, (**) $p < 0.01$, (***) $p < 0.001$, (****) $p < 0.0001$.
1084

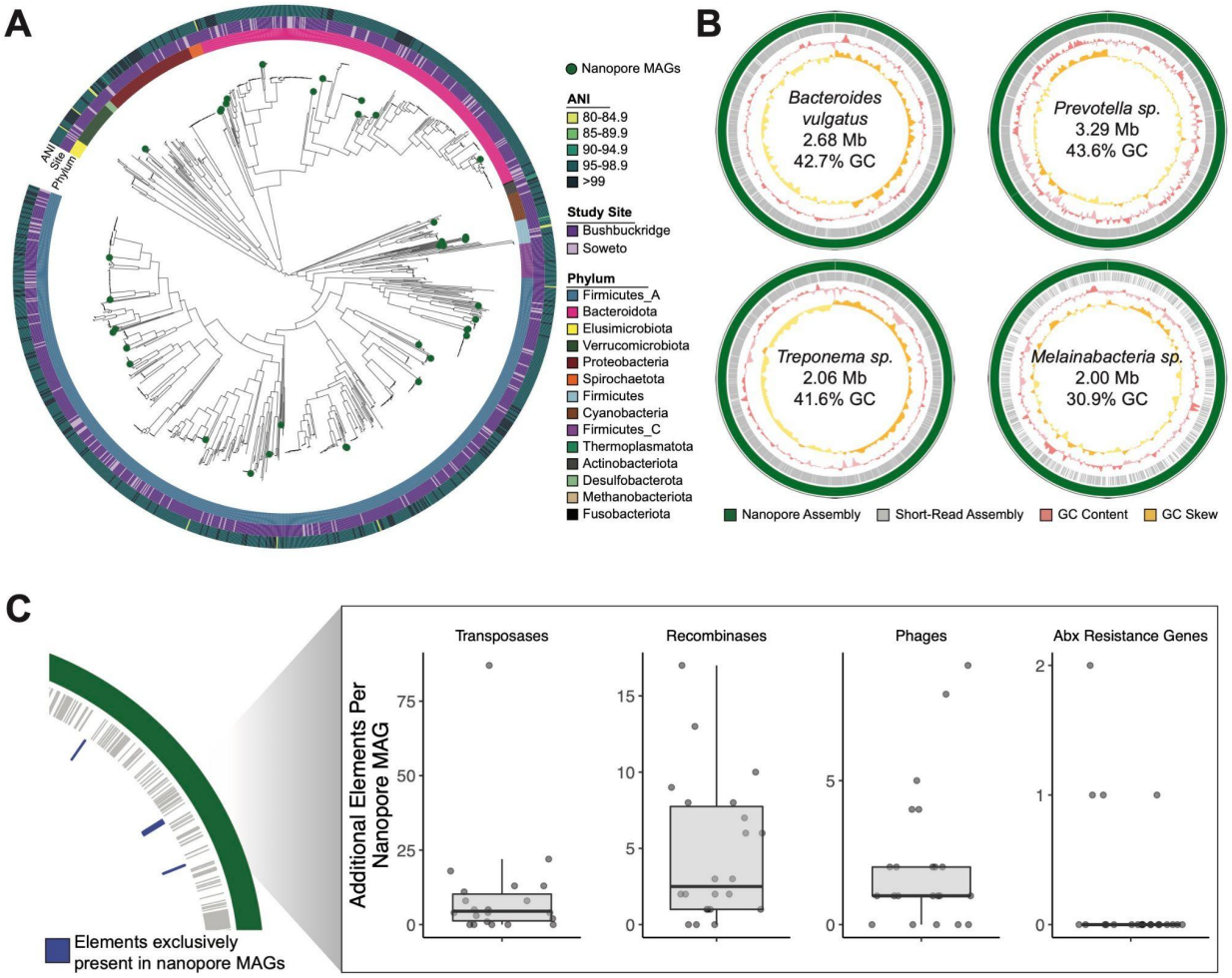


Figure 5. Complete and contiguous genomes of South African microbiota

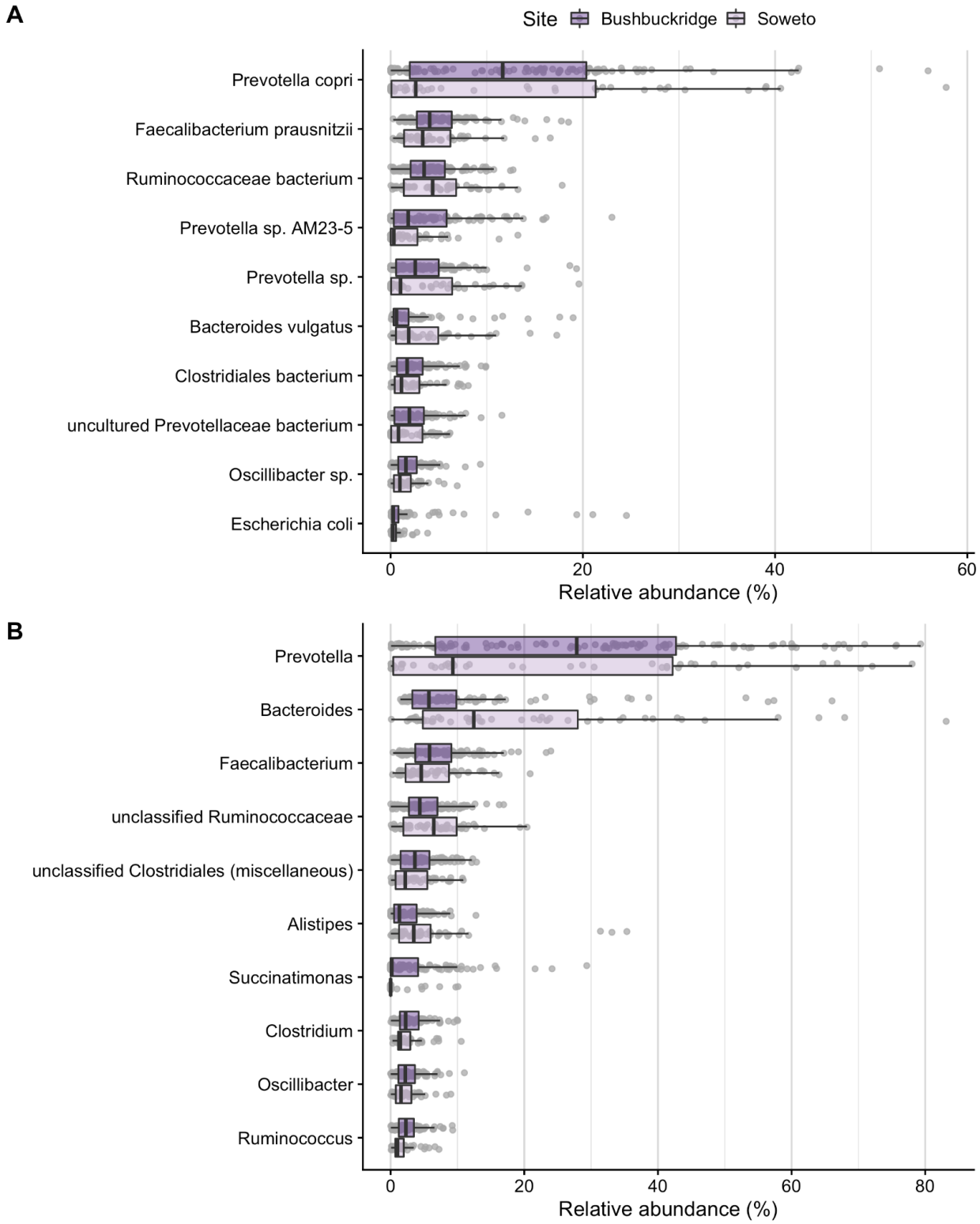
(A) Phylogenetic tree of dereplicated short-read MAGs and medium- and high-quality nanopore MAGs (green circles). Innermost ring indicates GTDB phylum, middle ring indicates study site associated with each MAG, and outer ring indicates the highest average nucleotide identity between each MAG and genomes from UHGG.

(B) A selection of MAGs assembled from long-read sequencing (green) of three South African samples compared contigs assembled from corresponding short read data (grey). Third track (pink) indicates sliding genomic GC content, and fourth track (yellow) indicates sliding genomic GC skew. Breaks in circles represent different contigs. Genomic information within plots refer to assembly statistics of nanopore MAGs.

(C) Number of additional genomic elements present in medium- and high-quality nanopore MAGs that are absent in corresponding short-read MAGs for the same organism, as diagrammed in the left hand panel. Box plot lower and upper hinges

1099 correspond to the first and third quartiles, upper and lower box plot whiskers represent
1100 the highest and lowest values within 1.5 times the interquartile range, and the
1101 horizontal line represents the median.
1102 (B) Taxonomic of de-replicated medium- and high-quality nanopore MAGs. Black
1103 circles represent nanopore MAGs, at the highest level of taxonomic classification by
1104 GTDB.
1105

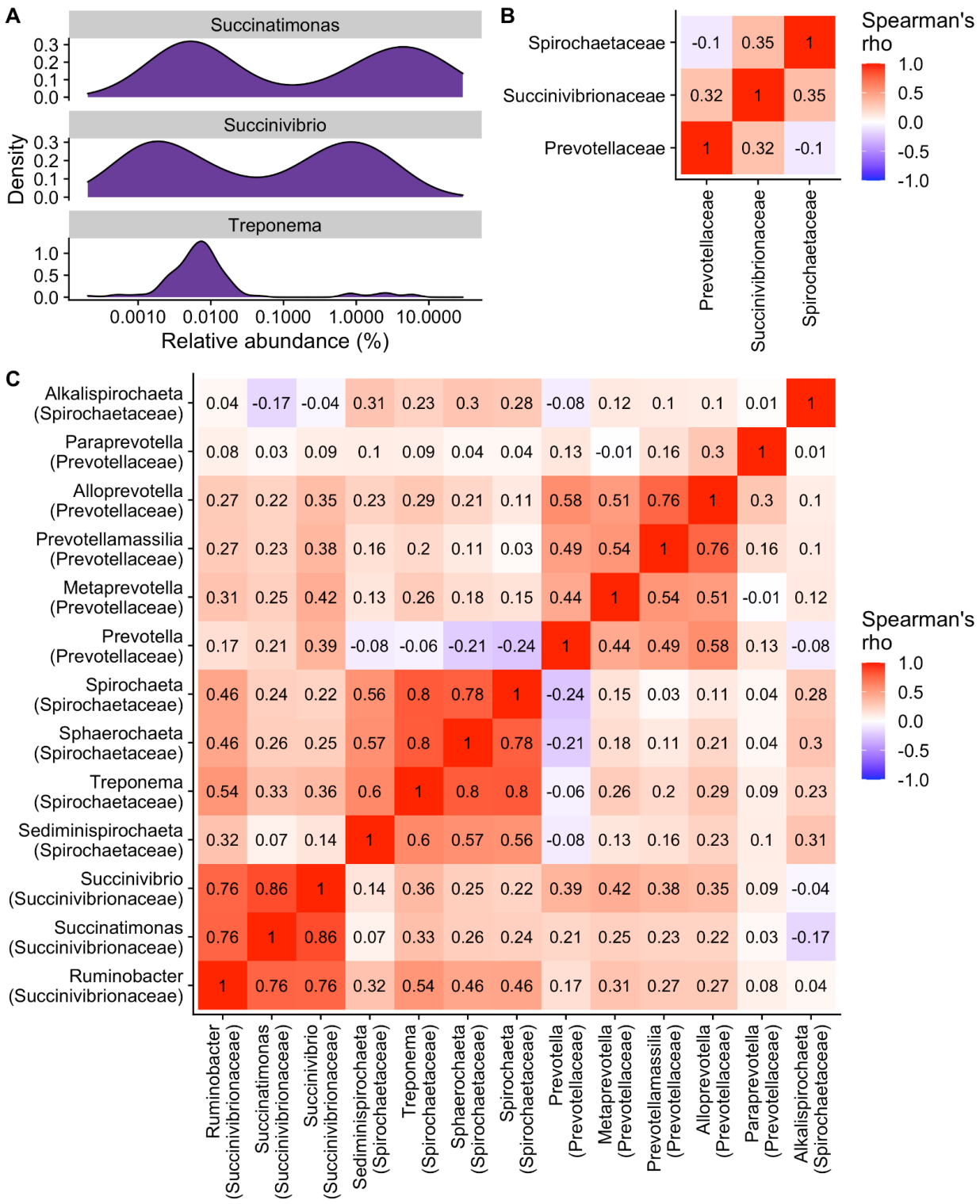
1106 # Supplementary Figures



1107

1108 **Supplementary Figure 1. Most abundant species and genera**

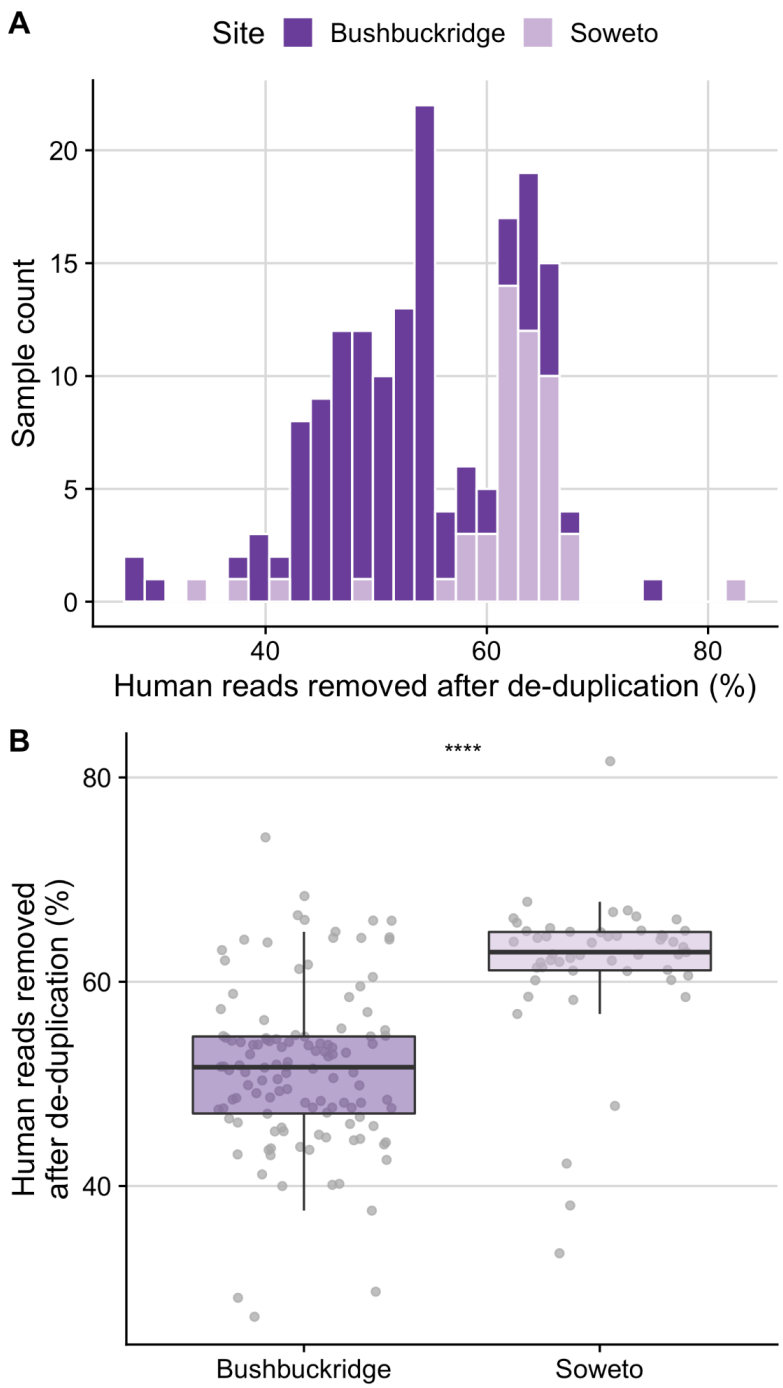
1109 Most abundant taxa by mean relative abundance (total sum scaling) shown for samples
1110 from Bushbuckridge (n=117) and Soweto (n=51). Taxa are plotted in decreasing order
1111 of mean relative abundance (vertical line) calculated across both cohorts combined.
1112 Lower and upper box plot hinges correspond to the first and third quartiles, upper and
1113 lower box plot whiskers represent the highest and lowest values within 1.5 times the
1114 interquartile range, and the vertical line represents the median.
1115 (A) The most abundant species are *Prevotella copri*, *Faecalibacterium prausnitzii,* and a
1116 bacterium from the family Ruminococcaceae.
1117 (B) *Prevotella*, *Bacteroides,* and *Faecalibacterium* are the most abundant genera
1118 across both 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.

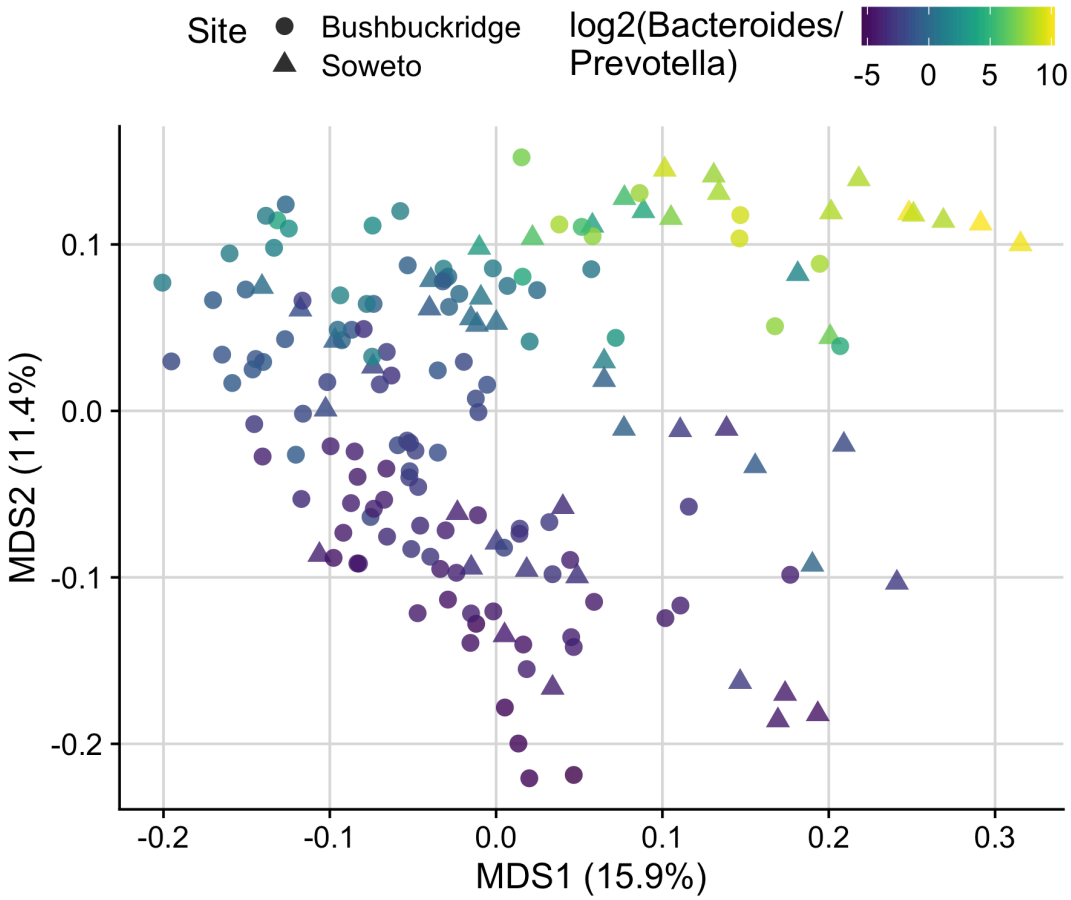
1123 Across all South African samples, several VANISH families (B) and genera (C) are
1124 correlated, with the exception of *Prevotella* and genera of the family *Spirochaetaceae*
1125 which are not correlated with *Prevotella* (*Treponema*) or weakly anti-correlated with
1126 *Prevotella* (*Spirochaeta*, *Sphaerochaeta*, *Sediminispirochaeta*).



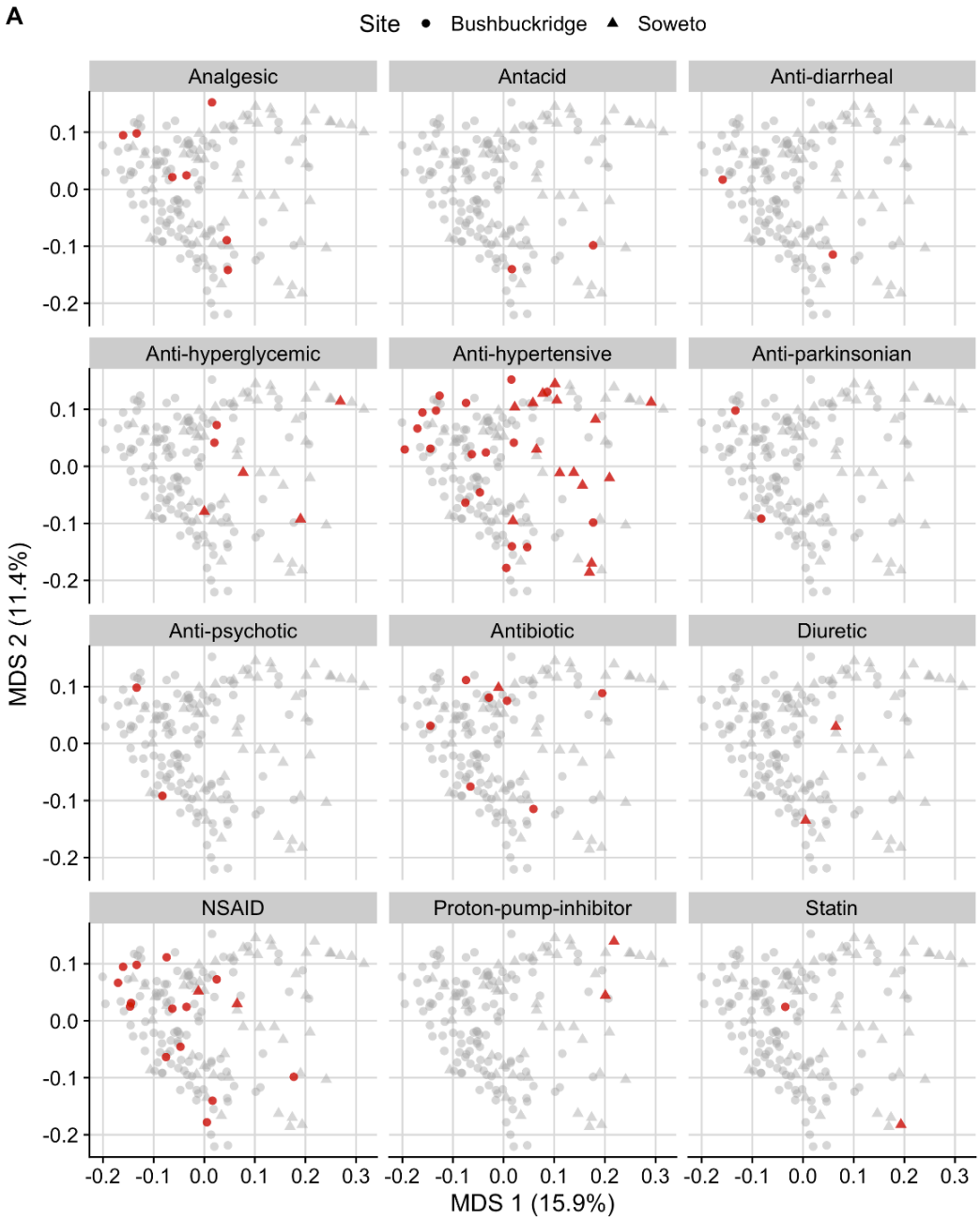
1127

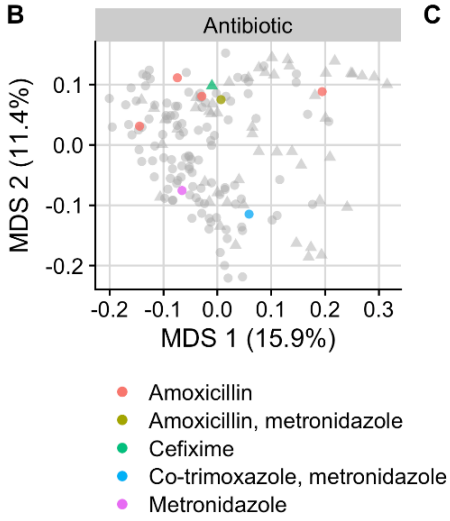
1128 **Supplementary Figure 3. Abundance of human reads in metagenomic sequencing**
1129 (A) Histogram and (B) box plots indicating that the proportion of human reads removed
1130 after deduplication was found to be higher in the Soweto cohort compared to
1131 Bushbuckridge (Two-sided Wilcoxon rank sum test, p = 1.661e-12). Significance
1132 values for Wilcoxon rank sum tests denoted as (****) for $p < 0.0001$. Lower and upper
1133 box plot hinges correspond to the first and third quartiles, upper and lower box plot

whiskers represent the highest and lowest values within 1.5 times the interquartile range, and the vertical line represents the median.



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 (CSS normalized) colored by log2 ratio of the relative abundance of the genera *Bacteroides* and *Prevotella*. *Bacteroides* and *Prevotella* are major axes of variation across study samples.





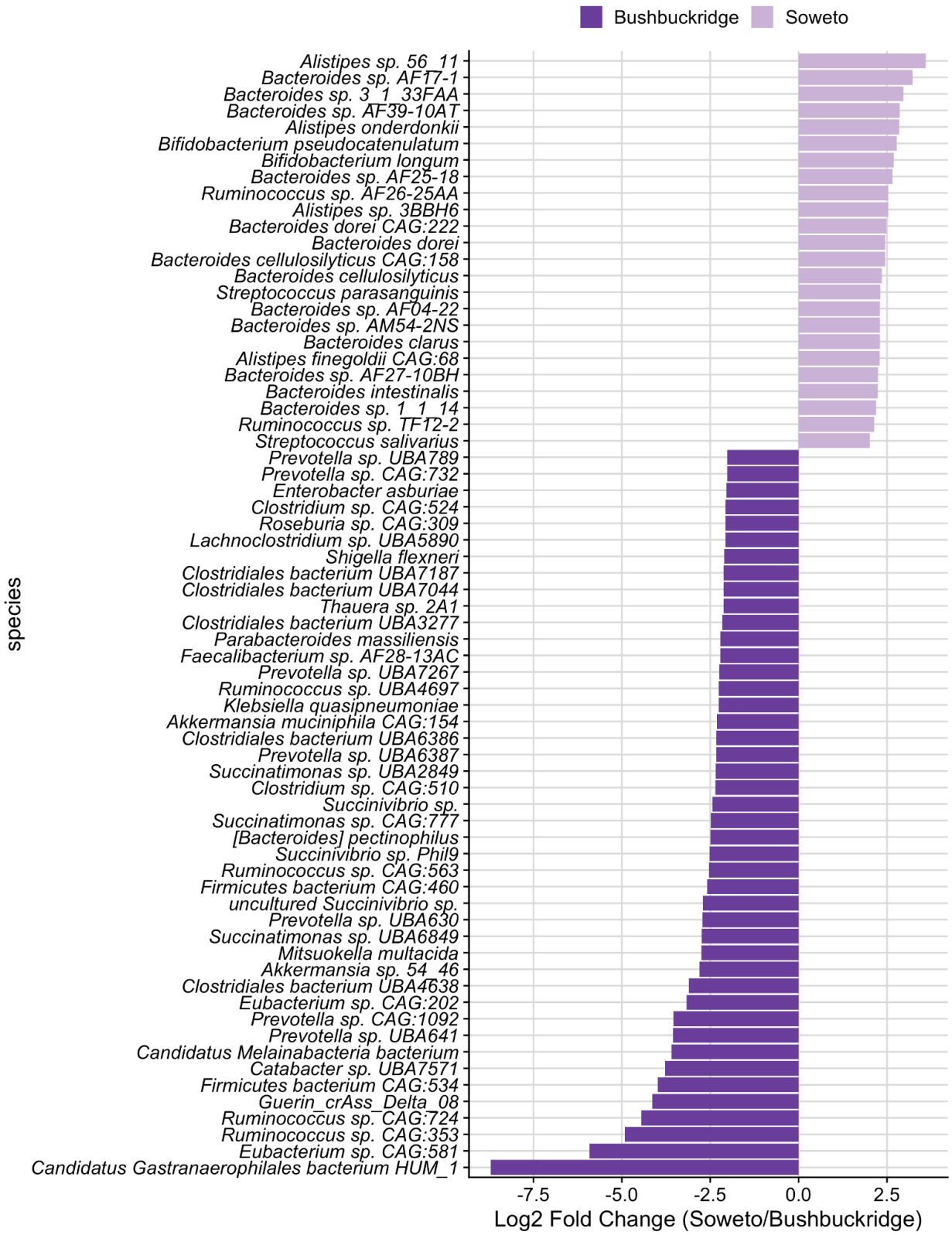
C

Category	R2	Pr(>F)	FDR
Analgesic	0.005	0.644	0.781
Antacid	0.006	0.367	0.781
Anti-diarrheal	0.005	0.468	0.781
Anti-hyperglycemic	0.010	0.041	0.246
Anti-hypertensive	0.005	0.781	0.781
Anti-parkinsonian	0.005	0.660	0.781
Anti-psychotic	0.005	0.651	0.781
Antibiotic	0.005	0.758	0.781
Diuretic	0.006	0.318	0.781
NSAID	0.007	0.247	0.781
Proton-pump-inhibitor	0.013	0.036	0.246
Statin	0.006	0.402	0.781

1145

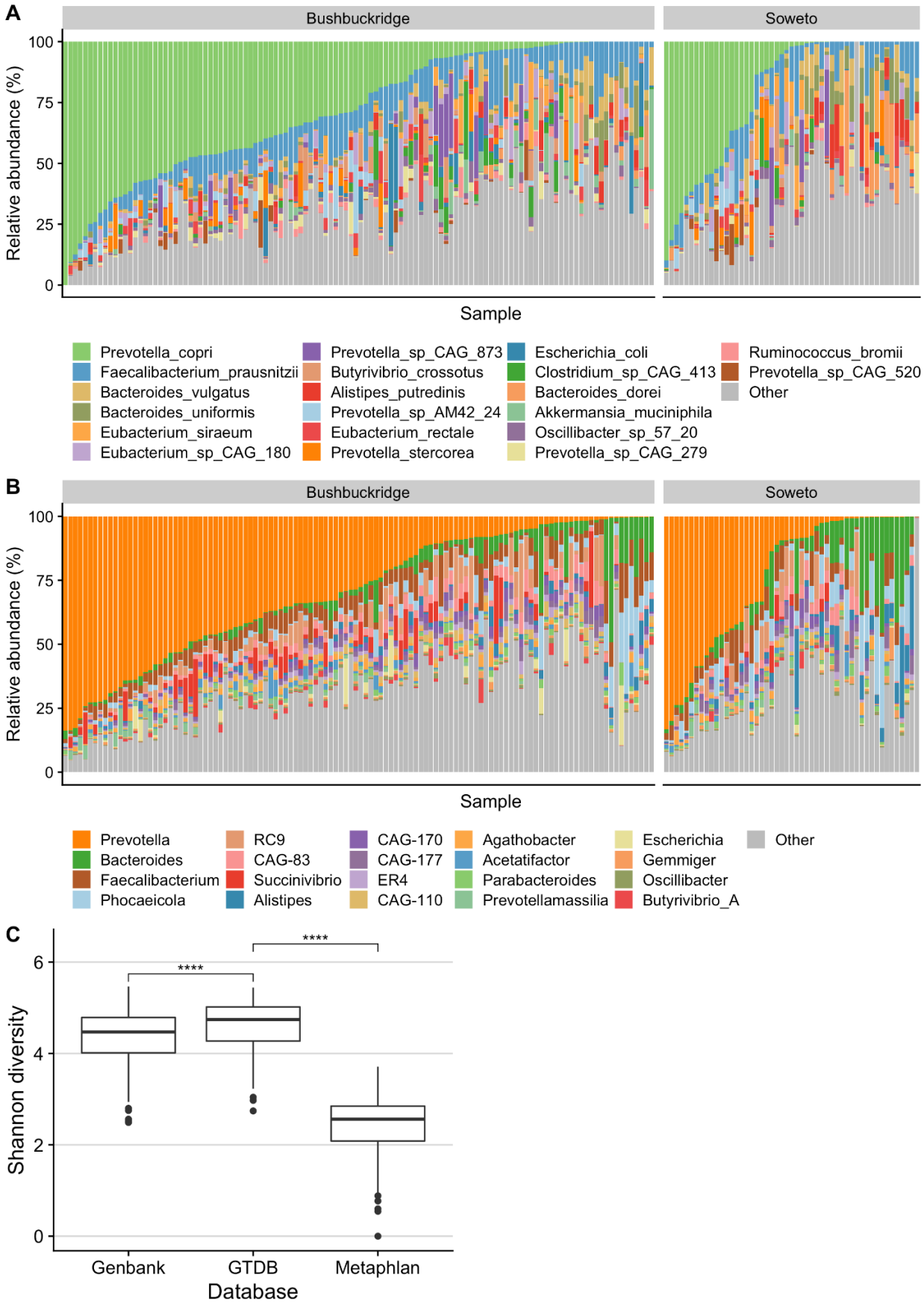
1146 **Supplementary Figure 5: Concomitant medications do not substantially impact**
1147 **community composition**
1148 Multidimensional scaling ordination of Bray-Curtis distance calculated from species
1149 classifications. Circles indicate participants from Bushbuckridge, triangles indicate
1150 participants from Soweto.
1151 (A) Points are colored red if the participant was taking a medication of the
1152 corresponding class, patients not taking a medication of that class are shown in gray.
1153 (B) Specific antibiotics taken by participants. Points are colored according to the
1154 antibiotic or combination of antibiotics reported.
1155 (C) PERMANOVA R^2 values and nominal and adjusted p-values for the variation
1156 explained by each drug class. Pr(>F) is the unadjusted p-value associated with the
1157 PERMANOVA F statistic, and FDR is the adjusted p-value to control the false discovery
1158 rate.
1159

1160


Bushbuckridge
Soweto
species
Alistipes sp. 56_11
Bacteroides sp. AF17-1
Bacteroides sp. 3_1_33FAA
Bacteroides sp. AF39-10AT
Alistipes onderdonkii
Bifidobacterium pseudocatenulatum
Bifidobacterium longum
Bacteroides sp. AF25-18
Ruminococcus sp. AF26-25AA
Alistipes sp. 3BBH6
Bacteroides dorei CAG:222
Bacteroides dorei
Bacteroides cellulosilyticus CAG:158
Bacteroides cellulosilyticus
Streptococcus parasanguinis
Bacteroides sp. AF04-22
Bacteroides sp. AM54-2NS
Bacteroides clarus
Alistipes finegoldii CAG:68
Bacteroides sp. AF27-10BH
Bacteroides intestinalis
Bacteroides sp. 1_1_14
Ruminococcus sp. TF12-2
Streptococcus salivarius
Prevotella sp. UBA789
Prevotella sp. CAG:732
Enterobacter asburiae
Clostridium sp. CAG:524
Roseburia sp. CAG:309
Lachnoclostridium sp. UBA5890
Shigella flexneri
Clostridiales bacterium UBA7187
Clostridiales bacterium UBA7044
Thauera sp. 2A1
Clostridiales bacterium UBA3277
Parabacteroides massiliensis
Faecalibacterium sp. AF28-13AC
Prevotella sp. UBA7267
Ruminococcus sp. UBA4697
Klebsiella quasipneumoniae
Akkermansia muciniphila CAG:154
Clostridiales bacterium UBA6386
Prevotella sp. UBA6387
Succinatimonas sp. UBA2849
Clostridium sp. CAG:510
Succinivibrio sp.
Succinatimonas sp. CAG:777
[Bacteroides] pectinophilus
Succinivibrio sp. Phil9
Ruminococcus sp. CAG:563
Firmicutes bacterium CAG:460
uncultured Succinivibrio sp.
Prevotella sp. UBA630
Succinatimonas sp. UBA6849
Mitsuokella multacida
Akkermansia sp. 54_46
Clostridiales bacterium UBA4638
Eubacterium sp. CAG:202
Prevotella sp. CAG:1092
Prevotella sp. UBA641
Candidatus Melainabacteria bacterium
Catabacter sp. UBA7571
Firmicutes bacterium CAG:534
Guerin_crAss_Delta_08
Ruminococcus sp. CAG:724
Ruminococcus sp. CAG:353
Eubacterium sp. CAG:581
Candidatus Gastranaerophilales bacterium HUM_1
-7.5
-5.0
-2.5
0.0
2.5
Log2 Fold Change (Soweto/Bushbuckridge)

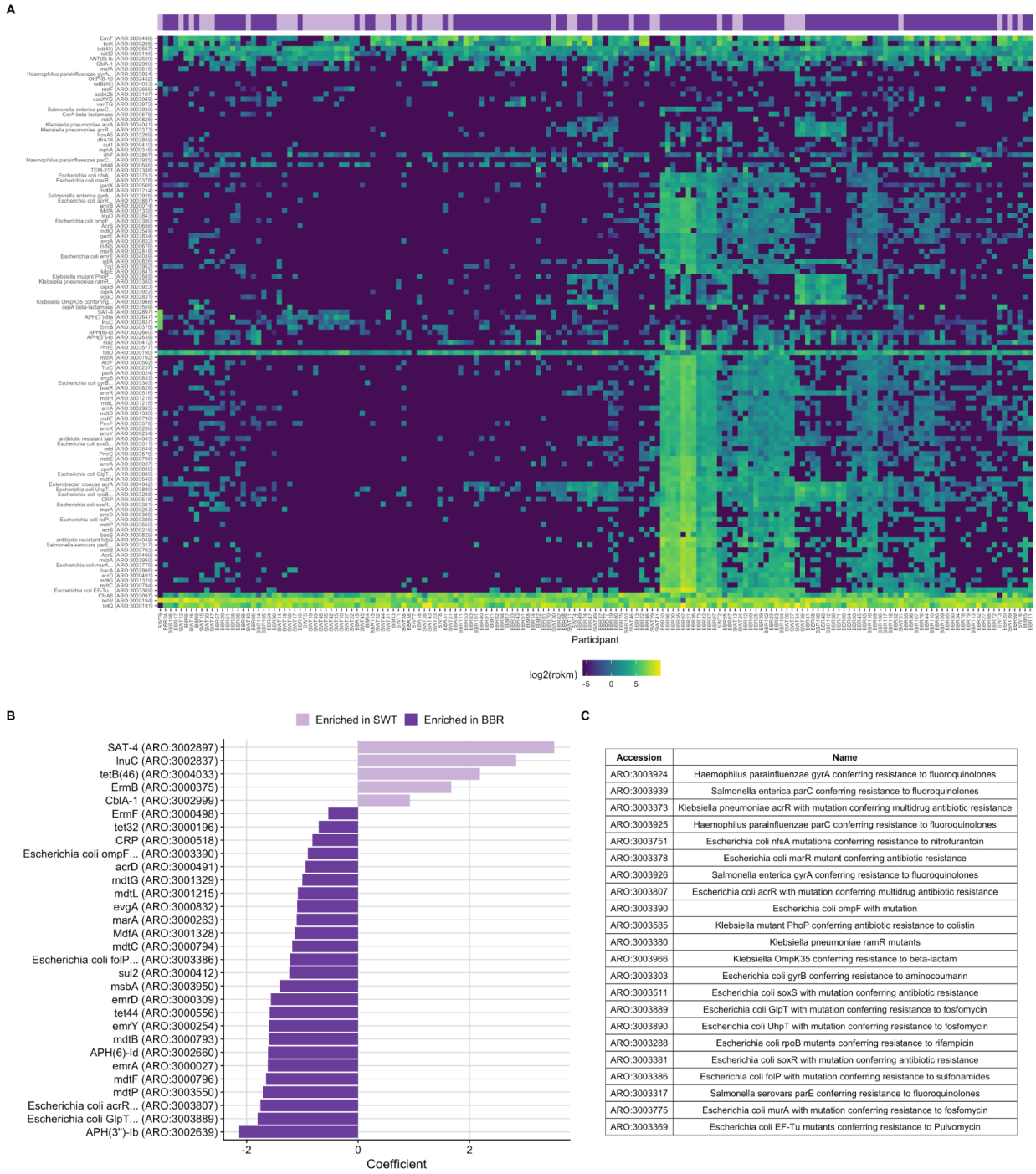
1161

1162 **Supplementary Figure 6. Differentially abundant species between Bushbuckridge**
1163 **and Soweto**
1164 Differentially abundant microbial species between rural Bushbuckridge and urban
1165 Soweto samples identified by DESeq2. Features with log2 fold change greater than
1166 one are shown (full results in Supplementary Table 7). Note that differentially abundant
1167 microbial genera are presented in Figure 2C.
1168
1169
1170


A
Bushbuckridge
Soweto
Relative abundance (%)
100
75
50
25
0
Sample
Prevotella_copri
Faecalibacterium_prausnitzii
Bacteroides_vulgatus
Bacteroides_uniformis
Eubacterium_siraeum
Eubacterium_sp_CAG_180
Prevotella_sp_CAG_873
Butyrivibrio_crossotus
Alistipes_putredinis
Prevotella_sp_AM42_24
Eubacterium_rectale
Prevotella_stercorea
Escherichia_coli
Clostridium_sp_CAG_413
Bacteroides_dorei
Akkermansia_muciniphila
Oscillibacter_sp_57_20
Prevotella_sp_CAG_279
Ruminococcus_bromii
Prevotella_sp_CAG_520
Other
B
Bushbuckridge
Soweto
Relative abundance (%)
100
75
50
25
0
Sample
Prevotella
Bacteroides
Faecalibacterium
Phocaeicola
RC9
CAG-83
Succinivibrio
Alistipes
CAG-170
CAG-177
ER4
CAG-110
Agathobacter
Acetatifactor
Parabacteroides
Prevotellamassilia
Escherichia
Gemmiger
Oscillibacter
Butyrivibrio_A
Other
C
Shannon diversity
6
4
2
0
****
****
Genbank
GTDB
Metaphlan
Database

1171

1172 **Supplementary Figure 7: GTDB yields increased taxonomic precision and alpha**
1173 **diversity**
1174 (A) Genus-level taxonomy using the MetaPhlAn3 classifier and database.
1175 (B) Genus-level taxonomy using the Genome Taxonomy Database (GTDB) release 95.
1176 (C) Shannon diversity across our custom GenBank database, the GTDB, and
1177 MetaPhlAn3. Shannon diversity is significantly higher using the GTDB as a reference
1178 collection compared to the custom GenBank database (Two-sided Wilcoxon rank sum
1179 test, $p = 4.929e-06$) and MetaPhlAn3 (Two-sided Wilcoxon rank sum test, $p < 2.2e-16$).
1180 Significance values for Wilcoxon rank sum tests denoted in the plot as (****) to
1181 represent $p < 0.0001$.
1182

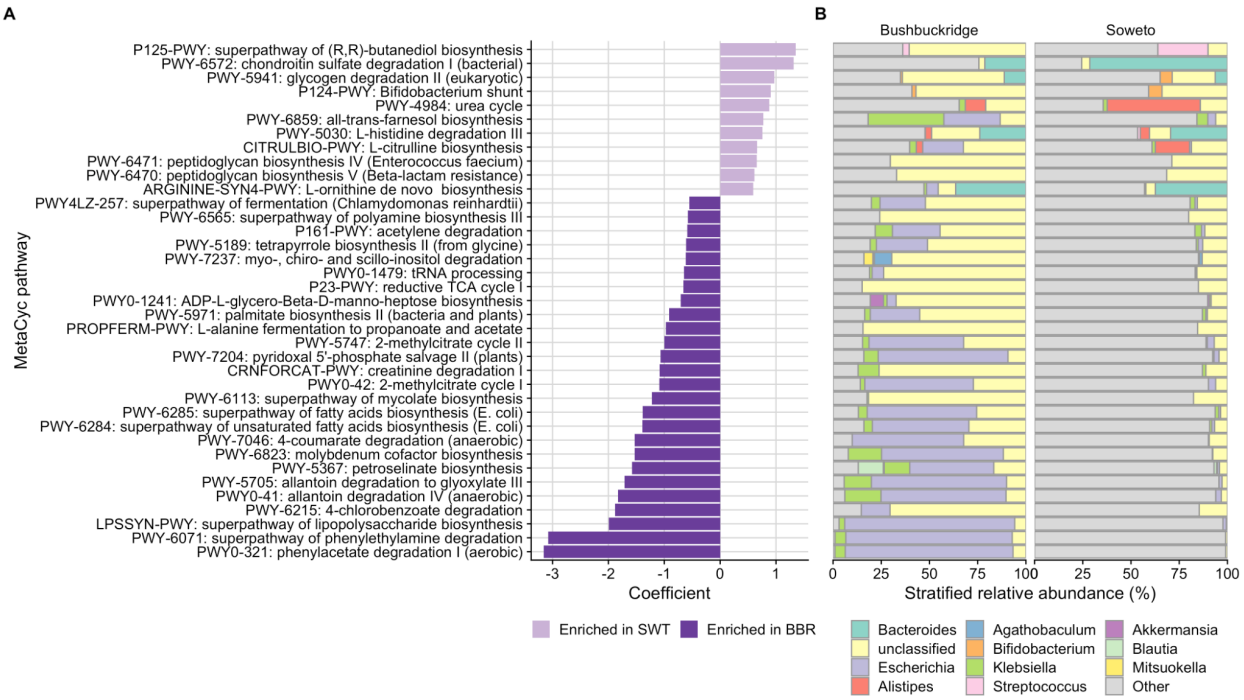


Accession	Name
ARO:3003924	Haemophilus parainfluenzae gyrA conferring resistance to fluoroquinolones
ARO:3003939	Salmonella enterica parC conferring resistance to fluoroquinolones
ARO:3003373	Klebsiella pneumoniae acrR with mutation conferring multidrug antibiotic resistance
ARO:3003925	Haemophilus parainfluenzae parC conferring resistance to fluoroquinolones
ARO:3003751	Escherichia coli nfsA mutations conferring resistance to nitrofurantoin
ARO:3003378	Escherichia coli marR mutant conferring antibiotic resistance
ARO:3003926	Salmonella enterica gyrA conferring resistance to fluoroquinolones
ARO:3003807	Escherichia coli acrR with mutation conferring multidrug antibiotic resistance
ARO:3003390	Escherichia coli ompF with mutation
ARO:3003585	Klebsiella mutant PhoP conferring antibiotic resistance to colistin
ARO:3003380	Klebsiella pneumoniae ramR mutants
ARO:3003966	Klebsiella OmpK35 conferring resistance to beta-lactam
ARO:3003303	Escherichia coli gyrB conferring resistance to aminocoumarin
ARO:3003511	Escherichia coli soxS with mutation conferring antibiotic resistance
ARO:3003889	Escherichia coli GlpT with mutation conferring resistance to fosfomycin
ARO:3003890	Escherichia coli UhpT with mutation conferring resistance to fosfomycin
ARO:3003288	Escherichia coli rpoB mutants conferring resistance to rifampicin
ARO:3003381	Escherichia coli soxR with mutation conferring antibiotic resistance
ARO:3003386	Escherichia coli folP with mutation conferring resistance to sulfonamides
ARO:3003317	Salmonella serovars parE conferring resistance to fluoroquinolones
ARO:3003775	Escherichia coli murA with mutation conferring resistance to fosfomycin
ARO:3003369	Escherichia coli EF-Tu mutants conferring resistance to Pulvomycin

Supplementary Figure 8: Differentially abundant antibiotic resistance genes between Bushbuckridge and Soweto

Antibiotic resistance genes were profiled using shortBRED against the Comprehensive Antibiotic Resistance Database (CARD). The shortBRED profiles were generated by grouping genes by CARD antibiotic resistance ontology (ARO) accession.

1189 (A) Heatmap showing log-transformed RPKM (reads per kilobase per million) values for
1190 antibiotic resistance genes in the gut metagenome of each participant. Columns
1191 (participants) are clustered by Canberra distance, rows (genes) are clustered by
1192 Euclidean distance.
1193 (B) Differentially abundant antibiotic resistance genes in Bushbuckridge (BBR) versus
1194 Soweto (SWT). RPKM profiles were compared between study sites using MaAsLin v2
1195 and p-values were adjusted to control the false discovery rate (FDR). Of 113 antibiotic
1196 resistance genes tested, 30 with $q < 0.05$ are shown.
1197 (C) Full CARD names for AROs whose names were truncated for plotting purposes in
1198 (A) and (B).

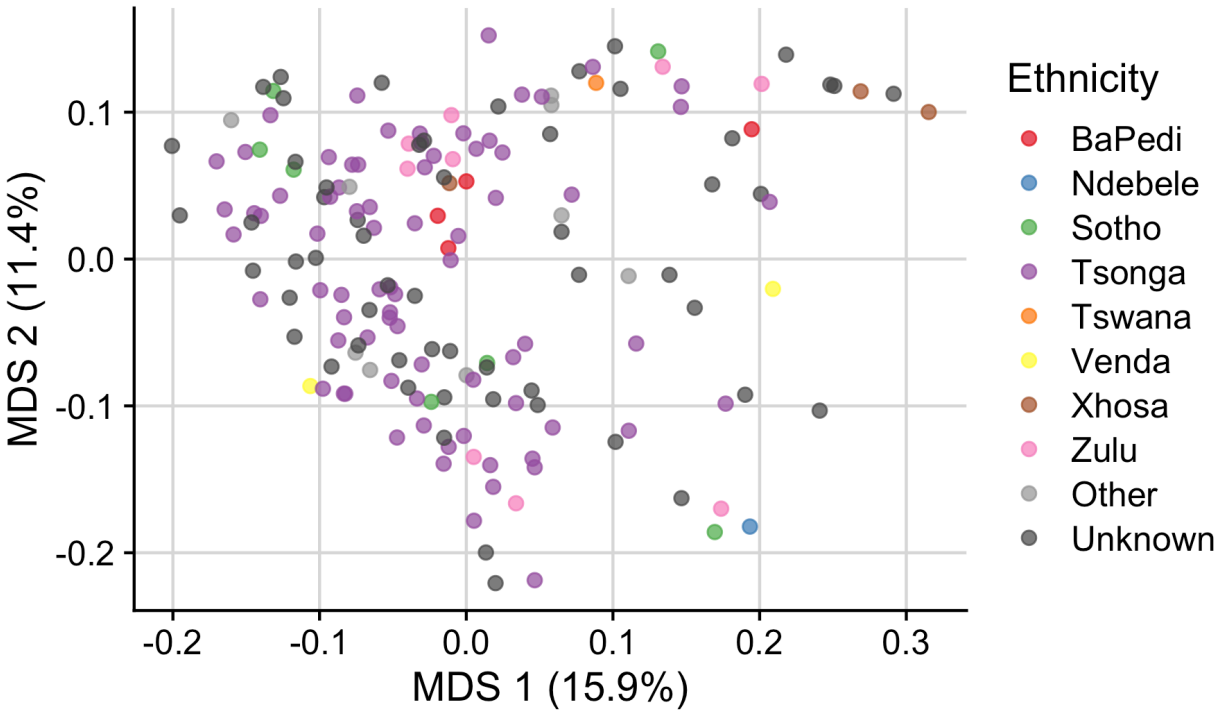


Supplementary Figure 9: Differential MetaCyc pathways between Bushbuckridge and Soweto

(A) MetaCyc pathways were profiled with HUMAnN v3 and differentially abundant pathways were identified using MaAsLin v2. 424 of 484 features (88%) met the 10% prevalence cutoff and 68 of 424 features (16%) were significantly differentially abundant between Bushbuckridge and Soweto with a q-value < 0.05. 37 features with q-value < 0.05 and coefficient >0.5 in either direction are shown.

(B) Stratified pathway composition by taxon for each significant MetaCyc pathway.

1209



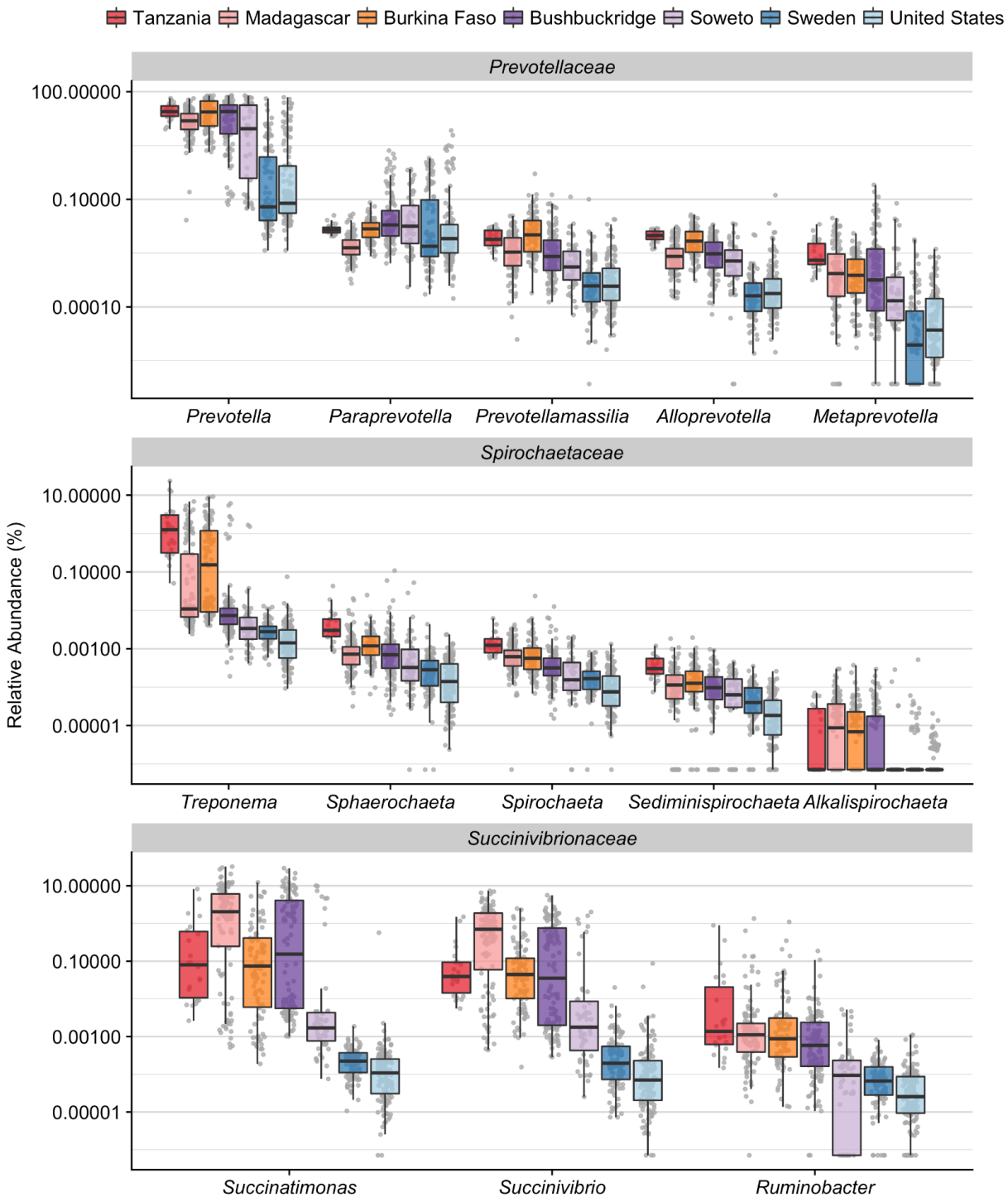
1210

1211 **Supplementary Figure 10. South African microbiomes do not cluster by self-**

1212 **reported ethnicity**

1213 Multidimensional scaling ordination of Bray-Curtis distance with samples are colored

1214 by self-reported ethnicity. Samples do not cluster by self-reported ethnicity.



1215

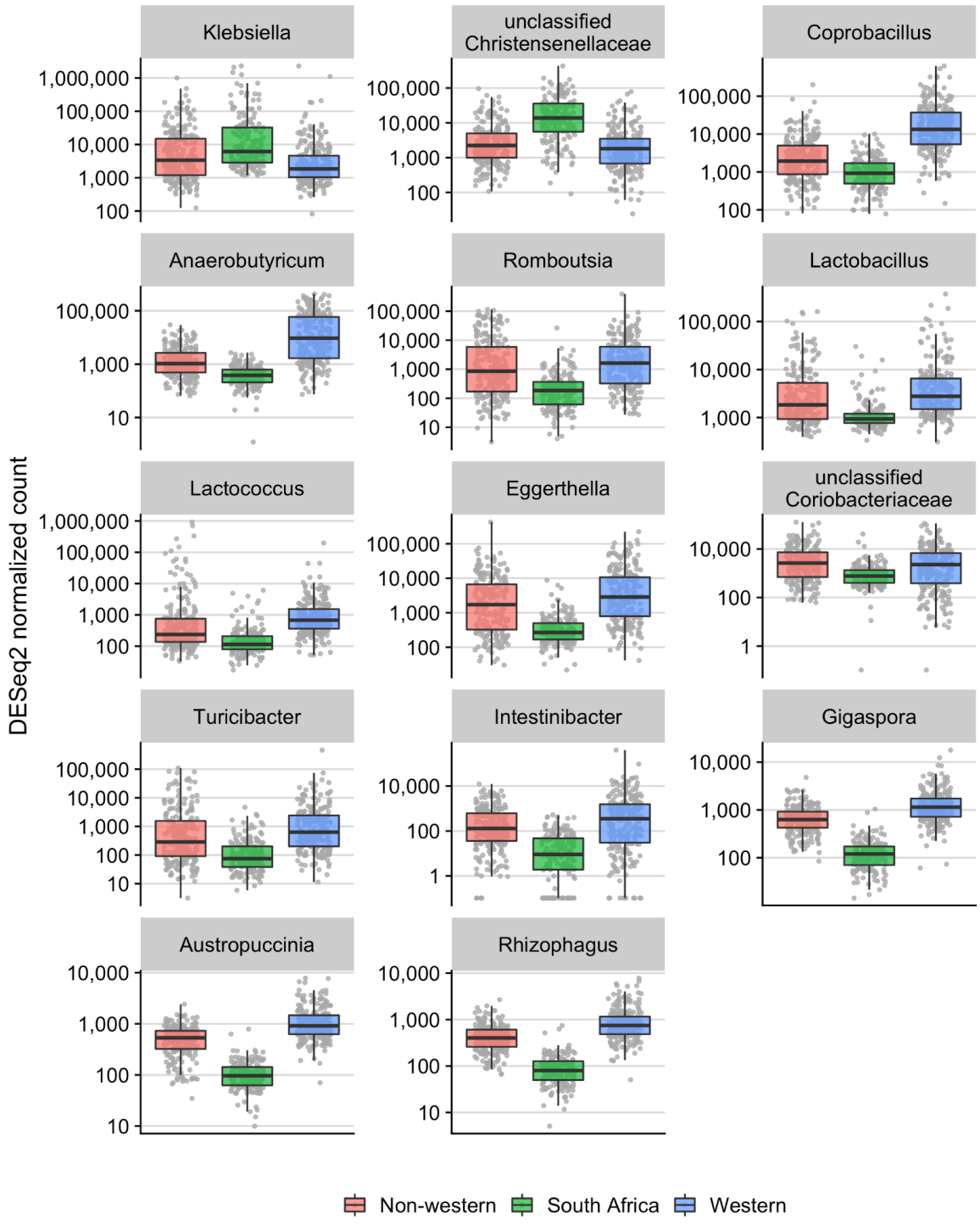
1216 **Supplementary Figure 11. Relative abundance of VANISH taxa in global cohort**

1217 Relative abundance of VANISH genera from the families Prevotellaceae,

1218 Spirochaetaceae, and Succinivibrionaceae. A pseudo-percent was substituted for zero

1219 values in order to plot on a log scale. Relative abundance values for most genera trend

toward decreasing from nonwestern cohorts to western cohorts. Box plot lower and upper hinges correspond to the first and third quartiles, upper and lower whiskers represent the highest and lowest values within 1.5 times the interquartile range, and the horizontal line represents the median.

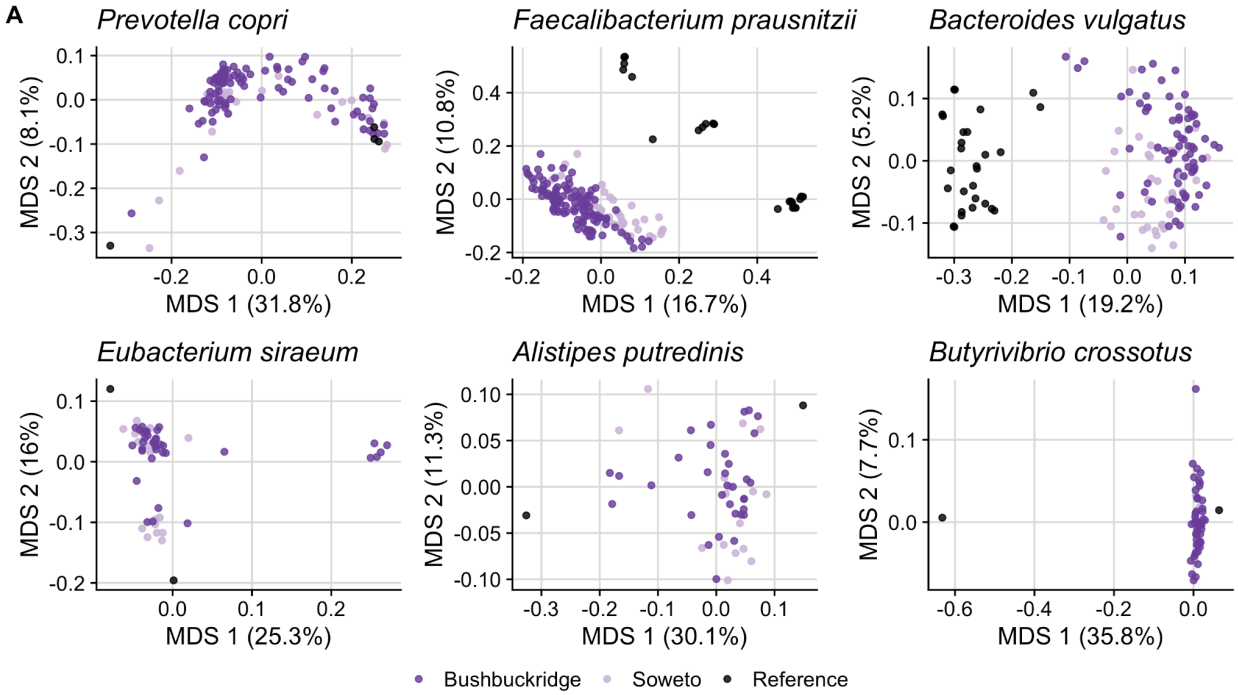


Supplementary Figure 12. Microbial genera enriched or depleted in South Africans relative to other cohorts

Samples were grouped by geographic region into “western” (USA, Sweden), “nonwestern” (Tanzania, Madagascar, Burkina Faso) and “South African” (Bushbuckridge, Soweto) and genera which distinguish the South African group from

1231 the western and nonwestern groups were determined using DESeq2. Genera present
1232 with at least 500 counts in 20% of samples were considered (190 features total). 14
1233 features with the same directionality of log2 fold change with respect to South Africa in
1234 both comparisons, with a minimum log2 fold change of 2 in each comparison, are
1235 shown. A pseudo-percent was added to zero values for plotting. Box plot lower and
1236 upper hinges correspond to the first and third quartiles, upper and lower whiskers
1237 represent the highest and lowest values within 1.5 times the interquartile range, and
1238 the horizontal line represents the median.
1239
1240

1241

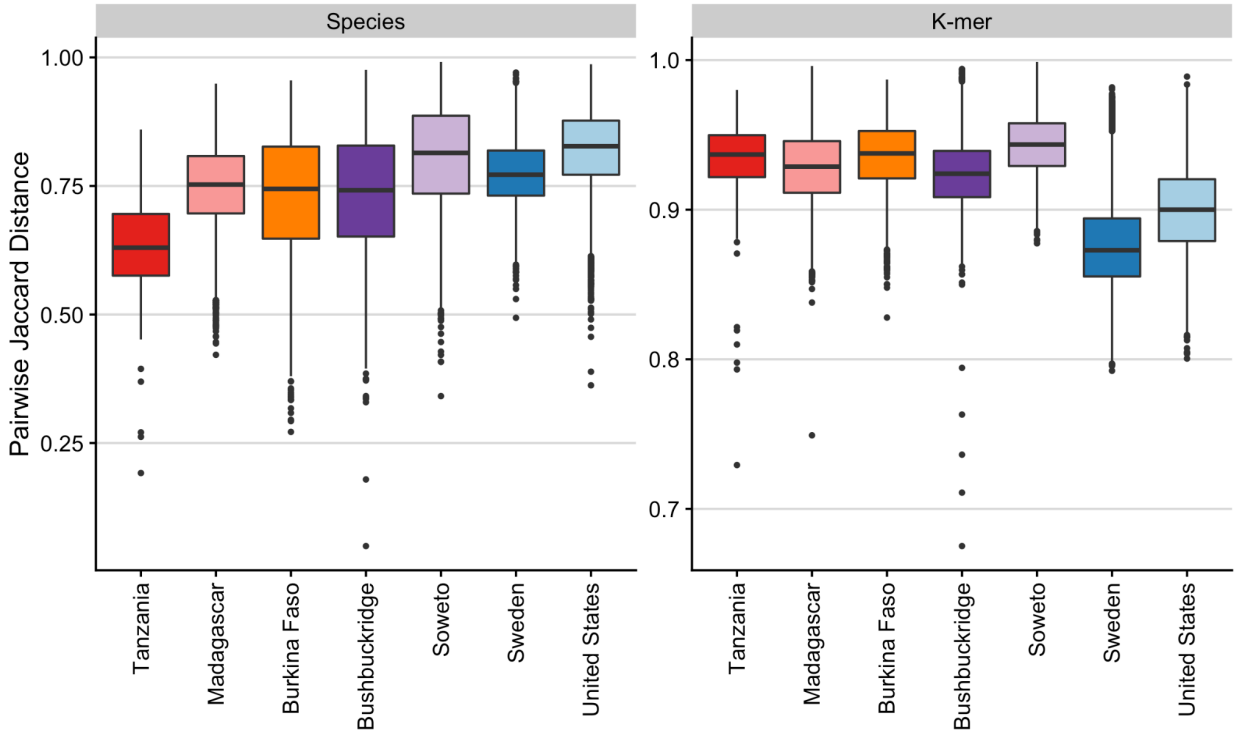


B

Species	R2	Pr(>F)	FDR
Prevotella copri	0.016	0.039	0.0585
Faecalibacterium prausnitzii	0.052	0.001	0.0030
Bacteroides vulgatus	0.027	0.001	0.0030
Eubacterium siraeum	0.046	0.011	0.0220
Alistipes putredinis	0.030	0.099	0.1188
Butyrivibrio crossotus	0.020	0.308	0.3080

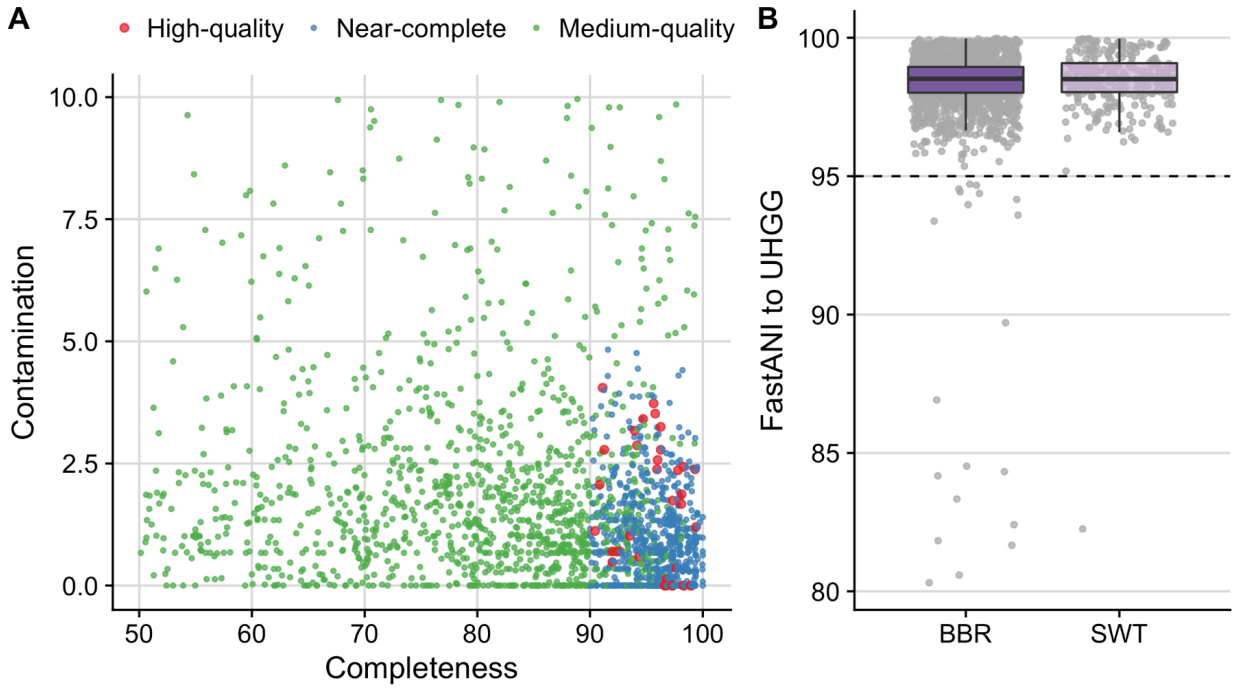
1242

1243 **Supplementary Figure 13: Pangenomes of South African metagenomic strains**
1244 (A) Multidimensional scaling (MDS) plots of Jaccard distance between pangenome
1245 content of the six most abundant bacteria cohort-wide as measured by MetaPhlAn3.
1246 (B) PERMANOVA results testing the null hypothesis that the centroids of
1247 Bushbuckridge and Soweto sample pangenomes differ in location. PR(>F) signifies the
1248 unadjusted p-value for the F statistic and FDR signifies p-values adjusted to control the
1249 false discovery rate.



Supplementary Figure 14. 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. Box plot lower and upper hinges correspond to the first and third quartiles, upper and lower whiskers represent the highest and lowest values within 1.5 times the interquartile range, and the horizontal line represents the median.



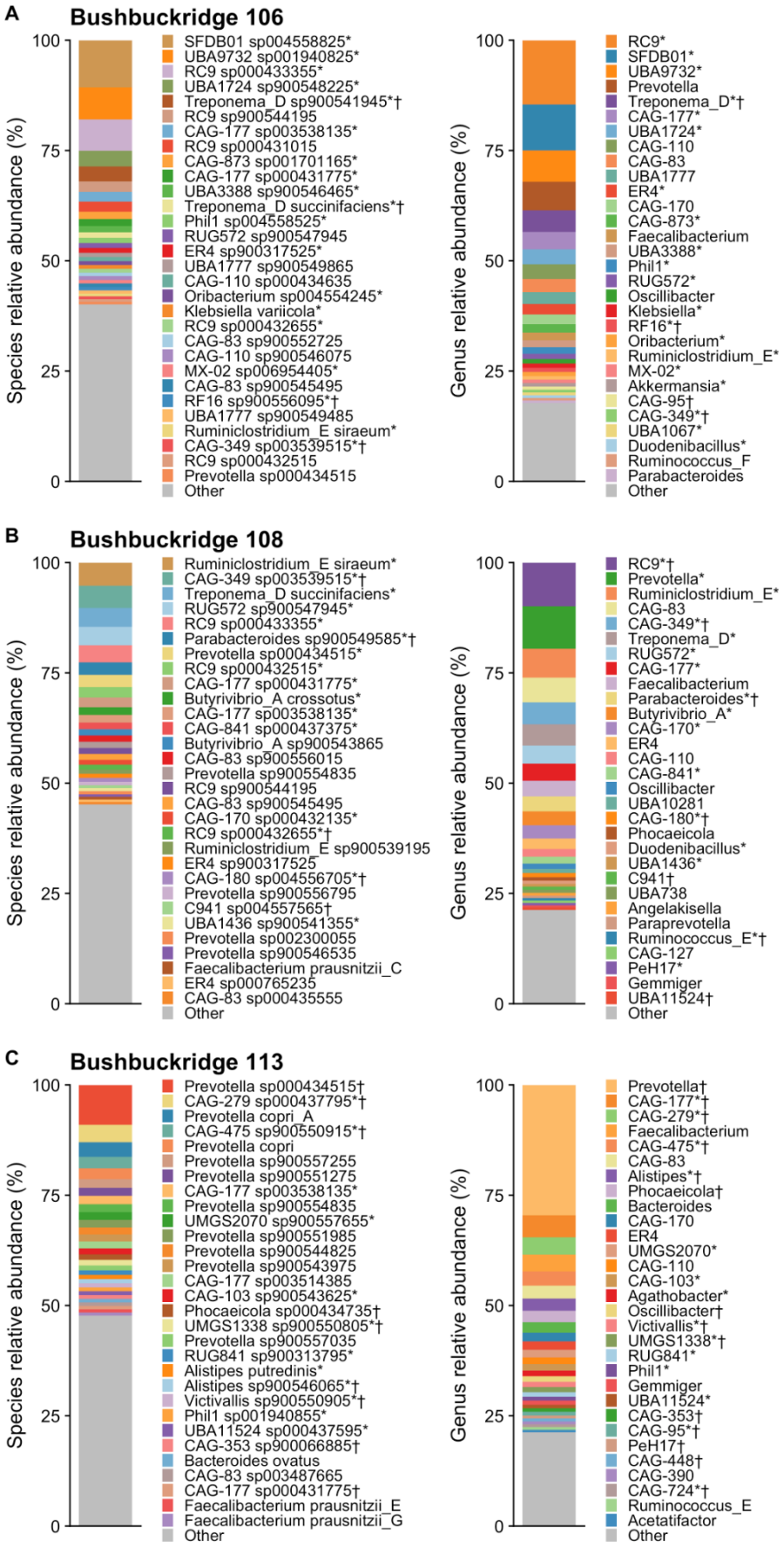
C

Participant	Phylum	Class	Order	Family	Genus	Species
BBR76	Actinobacteriota	Coriobacteriia	Coriobacteriales	Coriobacteriaceae	Collinsella	
BBR89	Actinobacteriota	Coriobacteriia	Coriobacteriales	Coriobacteriaceae	Collinsella	
BBR96	Actinobacteriota	Coriobacteriia	Coriobacteriales	Coriobacteriaceae	Collinsella	
BBR54	Bacteroidota	Bacteroidia	Bacteroidales	P3	UBA3388	
BBR98	Desulfobacterota	Desulfovibrionia	Desulfovibrionales	Desulfovibrionaceae	Mailhella	Mailhella massiliensis
SWT9	Firmicutes	Bacilli	Acholeplasmatales	Anaeroplasmataceae		
BBR38	Firmicutes	Bacilli	RF39	CAG-611	CAG-877	
BBR111	Firmicutes_A	Clostridia	Lachnospirales	Lachnospiraceae		
BBR55	Firmicutes_A	Clostridia	Lachnospirales	Lachnospiraceae	Oribacterium	
BBR55	Firmicutes_A	Clostridia	Oscillospirales	Oscillospiraceae	CAG-110	
BBR106	Firmicutes_A	Clostridia	Peptostreptococcales	Anaerovoracaceae	UBA1191	
BBR118	Firmicutes_A	Clostridia	UMGS1810	UMGS1810		
BBR23	Proteobacteria	Alphaproteobacteria	RF32	CAG-239		
BBR106	Verrucomicrobiota	Kiritimatiellae	RFP12	UBA1067	UBA1067	
SWT19	Verrucomicrobiota	Verrucomicrobiae	Opitutales	CAG-312	QALA01	

Supplementary Figure 15: Novel short-read MAGs

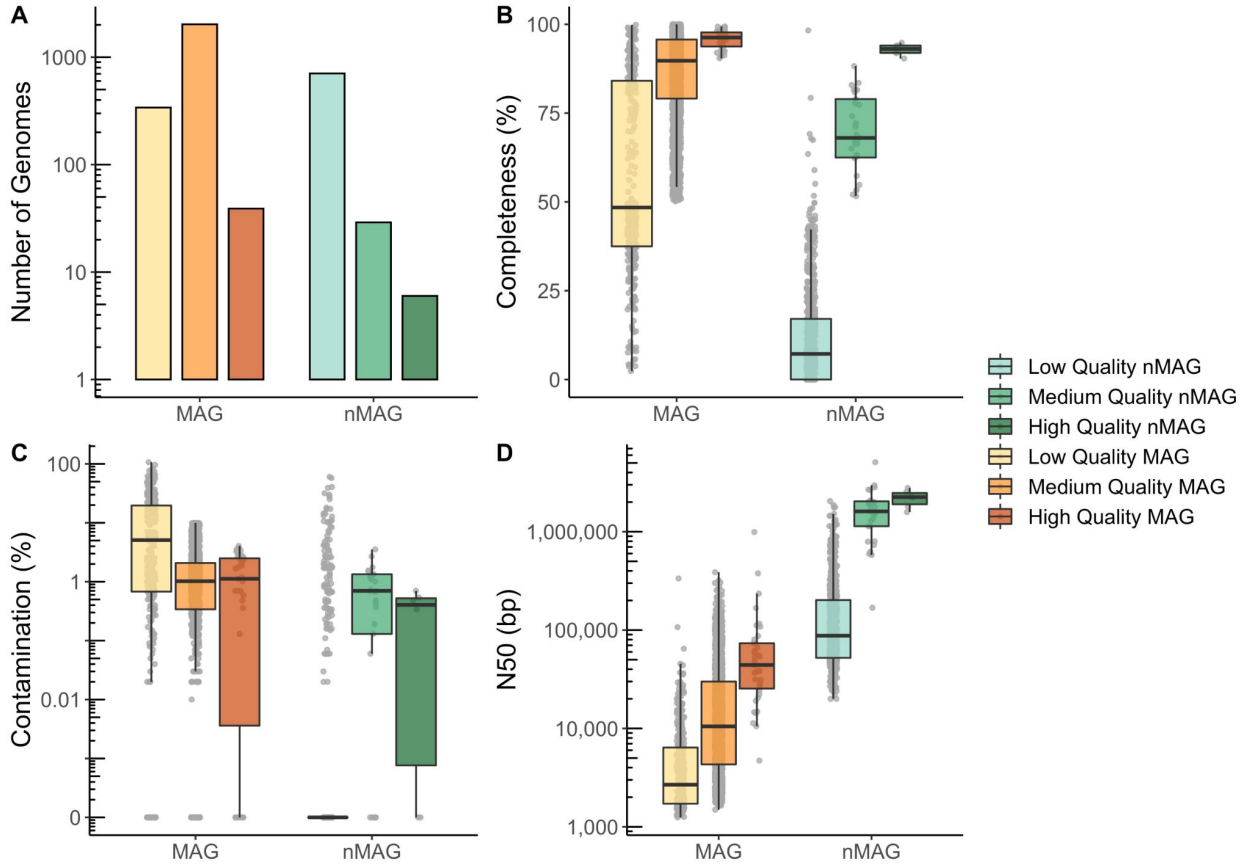
(A) Distribution of completeness and contamination (as assessed by CheckM software) in medium-quality (MQ), near-complete (NC), and high-quality (HQ) MAGs derived from Bushbuckridge (BBR) and Soweto (SWT). Smaller green points indicate MQ MAGs, smaller blue points indicate NC MAGs, and larger red dots indicate HQ MAGs. MQ

1266 MAGs must be >50% complete and <10% contaminated; NC MAGs must be ≥90%
1267 complete, ≤5% contaminated, and have a contig N50 ≥ 10 kb, average contig length ≥5
1268 kb, ≤500 contigs, and ≥90% of contigs with ≥5X read depth; HQ MAGs must be >90%
1269 complete, <5% contaminated, and have at least 18 tRNA genes and at least one each
1270 of the 5S, 16S, and 23S rRNA genes.
1271 (B) Distribution of FastANI average nucleotide identity values from each MQ or HQ
1272 MAG to the most closely related genome in the Unified Human Gastrointestinal
1273 Genome collection (UHGG). Not pictured are ten MQ MAGs with insufficient identity to
1274 any genome in UHGG such that FastANI could not be calculated. Box plot lower and
1275 upper hinges correspond to the first and third quartiles, upper and lower whiskers
1276 represent the highest and lowest values within 1.5 times the interquartile range, and
1277 the horizontal line represents the median.
1278 (C) Taxonomic classifications of “novel” MAGs from this study with <95% ANI to any
1279 genome in UHGG. Classifications according to GTDBtk using release 95 data.
1280


A
Bushbuckridge 106
Species relative abundance (%)
100
75
50
25
0
SFDB01 sp004558825*
UBA9732 sp001940825*
RC9 sp000433355*
UBA1724 sp900548225*
Treponema_D sp900541945*†
RC9 sp900544195
CAG-177 sp003538135*
RC9 sp000431015
CAG-873 sp001701165*
CAG-177 sp000431775*
UBA3388 sp900546465*
Treponema_D succinifaciens*†
Phil1 sp004558525*
RUG572 sp900547945
ER4 sp900317525*
UBA1777 sp900549865
CAG-110 sp000434635
Oribacterium sp004554245*
Klebsiella variicola*
RC9 sp000432655*
CAG-83 sp900552725
CAG-110 sp900546075
MX-02 sp006954405*
CAG-83 sp900545495
RF16 sp900556095*†
UBA1777 sp900549485
Ruminiclostridium_E siraeum*
CAG-349 sp003539515*†
RC9 sp000432515
Prevotella sp000434515
Other
Genus relative abundance (%)
RC9*
SFDB01*
UBA9732*
Prevotella
Treponema_D*†
CAG-177*
UBA1724*
CAG-110
CAG-83
UBA1777
ER4*
CAG-170
CAG-873*
Faecalibacterium
UBA3388*
Phil1*
RUG572*
Oscillibacter
Klebsiella*
RF16*†
Oribacterium*
Ruminiclostridium_E*
MX-02*
Akkermansia*
CAG-95†
CAG-349*†
UBA1067*
Duodenibacillus*
Ruminococcus_F
Parabacteroides
Other
B
Bushbuckridge 108
Species relative abundance (%)
Ruminiclostridium_E siraeum*
CAG-349 sp003539515*†
Treponema_D succinifaciens*
RUG572 sp900547945*
RC9 sp000433355*
Parabacteroides sp900549585*†
Prevotella sp000434515*
RC9 sp000432515*
CAG-177 sp000431775*
Butyrivibrio_A crossotus*
CAG-177 sp003538135*
CAG-841 sp000437375*
Butyrivibrio_A sp900543865
CAG-83 sp900556015
Prevotella sp900554835
RC9 sp900544195
CAG-83 sp900545495
CAG-170 sp000432135*
RC9 sp000432655*†
Ruminiclostridium_E sp900539195
ER4 sp900317525
CAG-180 sp004556705*†
Prevotella sp900556795
C941 sp004557565†
UBA1436 sp900541355*
Prevotella sp002300055
Prevotella sp900546535
Faecalibacterium prausnitzii_C
ER4 sp000765235
CAG-83 sp000435555
Other
Genus relative abundance (%)
RC9*†
Prevotella*
Ruminiclostridium_E*
CAG-83
CAG-349*†
Treponema_D*
RUG572*
CAG-177*
Faecalibacterium
Parabacteroides*†
Butyrivibrio_A*
CAG-170*
ER4
CAG-110
CAG-841*
Oscillibacter
UBA10281
CAG-180*†
Phocaeicola
Duodenibacillus*
UBA1436*
C941†
UBA738
Angelakisella
Paraprevotella
Ruminococcus_E*†
CAG-127
PeH17*
Gemmiger
UBA11524†
Other
C
Bushbuckridge 113
Species relative abundance (%)
Prevotella sp000434515†
CAG-279 sp000437795*†
Prevotella copri_A
CAG-475 sp900550915*†
Prevotella copri
Prevotella sp900557255
Prevotella sp900551275
CAG-177 sp003538135*
Prevotella sp900554835
UMGS2070 sp900557655*
Prevotella sp900551985
Prevotella sp900544825
Prevotella sp900543975
CAG-177 sp003514385
CAG-103 sp900543625*
Phocaeicola sp000434735†
UMGS1338 sp900550805*†
Prevotella sp900557035
RUG841 sp900313795*
Alistipes putredinis*
Alistipes sp900546065*†
Victivallis sp900550905*†
Phil1 sp001940855*
UBA11524 sp000437595*
CAG-353 sp900066885†
Bacteroides ovatus
CAG-83 sp003487665
CAG-177 sp000431775†
Faecalibacterium prausnitzii_E
Faecalibacterium prausnitzii_G
Other
Genus relative abundance (%)
Prevotella†
CAG-177*†
CAG-279*†
Faecalibacterium
CAG-475*†
CAG-83
Alistipes*†
Phocaeicola†
Bacteroides
CAG-170
ER4
UMGS2070*
CAG-110
CAG-103*
Agathobacter*
Oscillibacter†
Victivallis*†
UMGS1338*†
RUG841*
Phil1*
Gemmiger
UBA11524*
CAG-353†
CAG-95*†
PeH17†
CAG-448†
CAG-390
CAG-724*†
Ruminococcus_E
Acetatifactor
Other

1281

1282 **Supplementary Figure 16. Taxonomic composition for samples selected for**
1283 **nanopore sequencing**
1284 Short-read sequencing-based GTDB taxonomic classifications for the three samples
1285 selected for Nanopore sequencing, (A) Bushbuckridge 106, (B), Bushbuckridge 108, (C)
1286 Bushbuckridge 113. Species- and genus-level classifications shown for each sample.
1287 Top thirty taxa by relative abundance shown in each plot. Symbols indicate whether a
1288 medium- or high-quality short-read (*) or nanopore MAG (†) was assembled from the
1289 corresponding genus or species in the short read metagenomic data.
1290
1291



Supplementary Figure 17. Summary statistics for Illumina and nanopore MAGs generated from all samples.

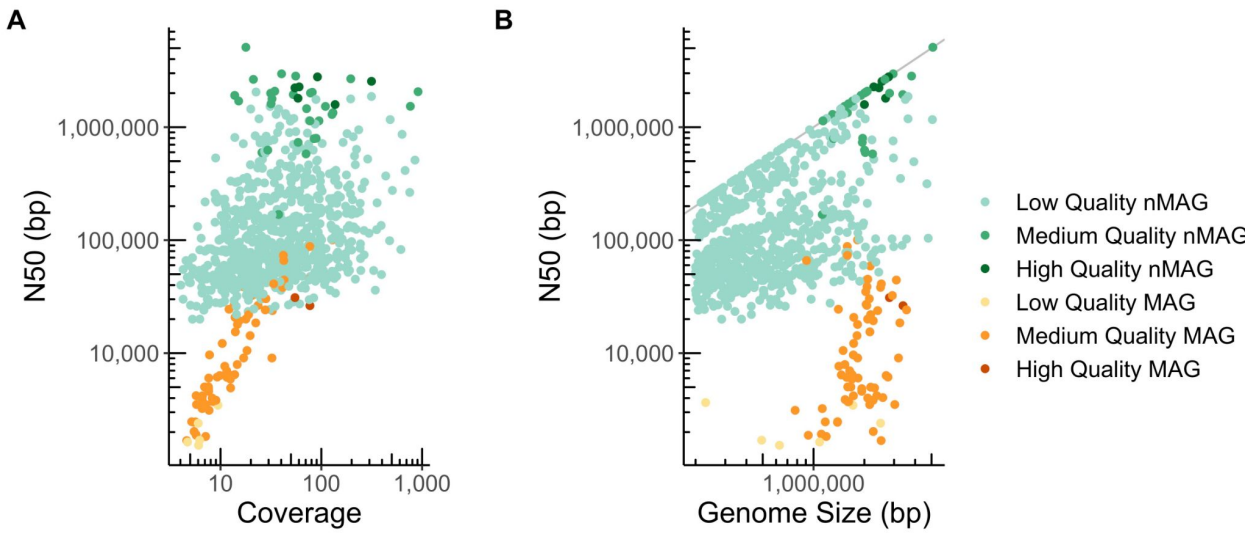
(A) Number of low-, medium-, and high-quality genomes as evaluated with Bowers et al. standards

(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.

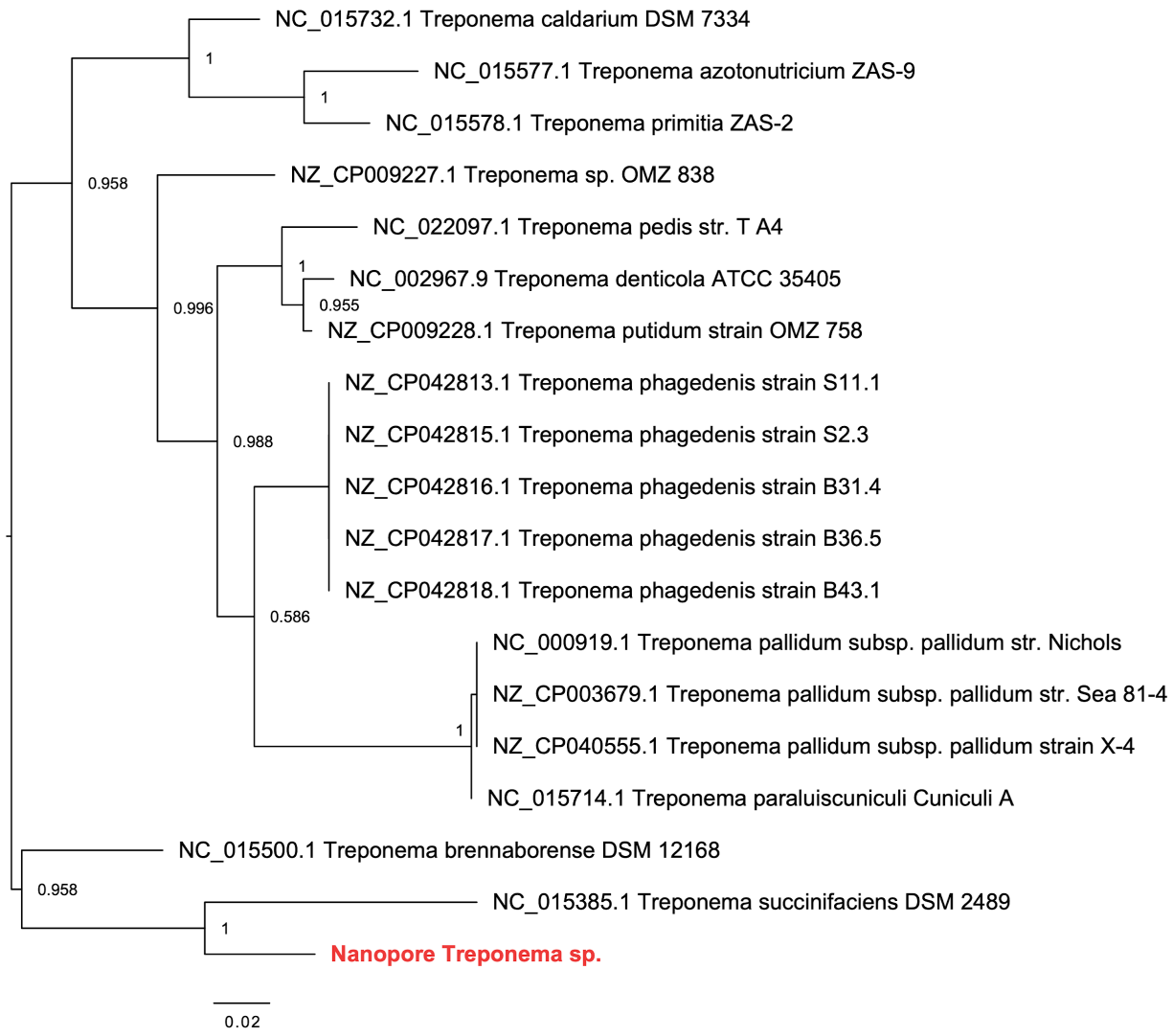
In all panels, box plot lower and upper hinges correspond to the first and third quartiles, upper and lower whiskers represent the highest and lowest values within 1.5 times the interquartile range, and the horizontal line represents the median.



Supplementary Figure 18. Summary statistics of nanopore and short read MAGs generated for three Bushbuckridge samples

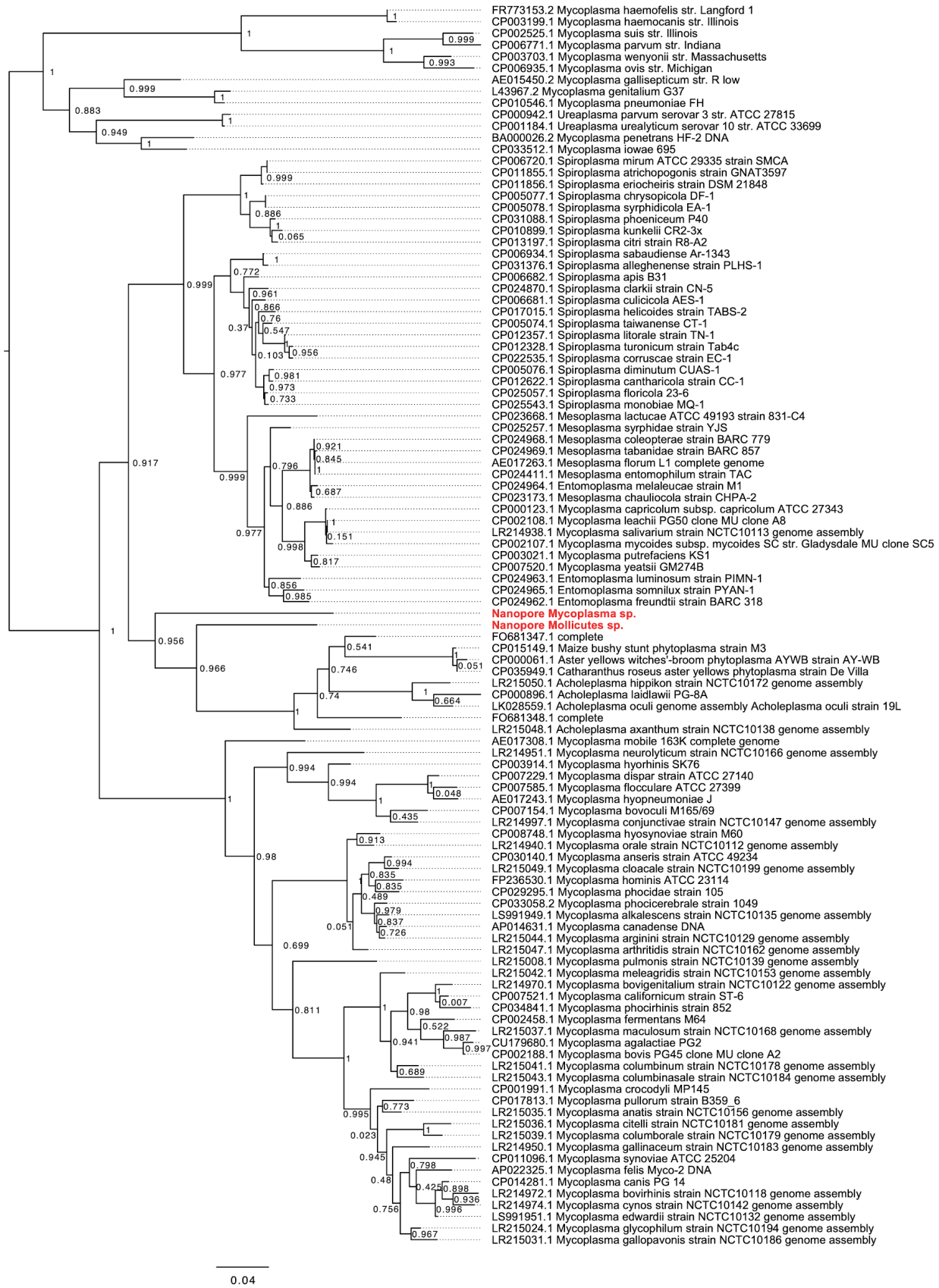
(A) MAG short read or long-read coverage versus MAG N50.

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



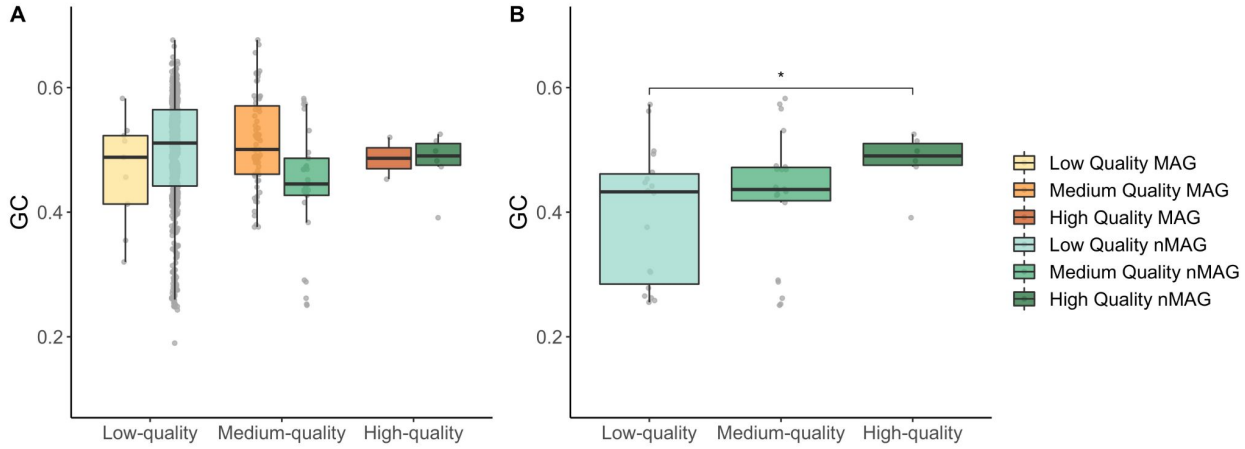
Supplementary Figure 19. Phylogeny of *Treponema* 16S rRNA sequences

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. Branch labels indicate Shimodaira-Hasegawa support values for splits.



Supplementary Figure 20. Phylogeny of Mollicutes 16S rRNA sequences

Phylogeny of 16S rRNA sequences from species of the class Mollicutes showing the Mollicutes and Mycoplasma genomes assembled via nanopore sequencing. Branch labels indicate Shimodaira-Hasegawa support values for splits.

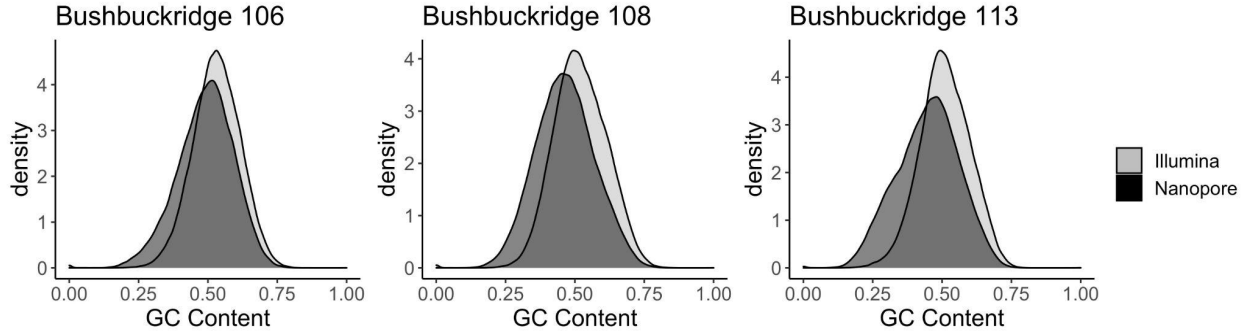


Supplementary Figure 21. GC content of MAGs and nMAGs generated from three Bushbuckridge samples

(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. (*) denotes $p \leq 0.05$, two-sided Wilcoxon rank sum test.

In both panels, box plot lower and upper hinges correspond to the first and third quartiles, upper and lower whiskers represent the highest and lowest values within 1.5 times the interquartile range, and the horizontal line represents the median.



1340 **Supplementary Figure 22. GC content of nanopore and Illumina sequencing reads**
1341 **generated from three Bushbuckridge samples**
1342 GC content was calculated for all processed Illumina reads (average length of 126 bp)
1343 and for 126 bp windows of all nanopore reads. GC content distribution was
1344 subsampled to 100,000 measurements per method.
1345

1346 # Main Text References

1347 1. Human Microbiome Project Consortium. Structure, function and diversity of the healthy
1348 human microbiome. *Nature* **486**, 207–214 (2012).
1349 2. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic
1350 sequencing. *Nature* **464**, 59–65 (2010).
1351 3. Gupta, V. K., Paul, S. & Dutta, C. Geography, Ethnicity or Subsistence-Specific Variations
1352 in Human Microbiome Composition and Diversity. *Front. Microbiol.* **8**, 1162 (2017).
1353 4. Brewster, R. *et al.* Surveying Gut Microbiome Research in Africans: Toward Improved
1354 Diversity and Representation. *Trends Microbiol.* (2019) doi:10.1016/j.tim.2019.05.006.
1355 5. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature*
1356 **486**, 222–227 (2012).
1357 6. Smits, S. A. *et al.* Seasonal cycling in the gut microbiome of the Hadza hunter-gatherers of
1358 Tanzania. *Science* **357**, 802–806 (2017).
1359 7. Rampelli, S. *et al.* Metagenome Sequencing of the Hadza Hunter-Gatherer Gut Microbiota.
1360 *Curr. Biol.* **25**, 1682–1693 (2015).
1361 8. Fragiadakis, G. K. *et al.* Links between environment, diet, and the hunter-gatherer
1362 microbiome. *Gut Microbes* **10**, 216–227 (2018).
1363 9. Schnorr, S. L. *et al.* Gut microbiome of the Hadza hunter-gatherers. *Nat. Commun.* **5**, 3654
1364 (2014).
1365 10. Hansen, M. E. B. *et al.* Population structure of human gut bacteria in a diverse cohort from
1366 rural Tanzania and Botswana. *Genome Biol.* **20**, 16 (2019).
1367 11. Obregon-Tito, A. J. *et al.* Subsistence strategies in traditional societies distinguish gut
1368 microbiomes. *Nat. Commun.* **6**, 6505 (2015).
1369 12. Pasolli, E. *et al.* Extensive Unexplored Human Microbiome Diversity Revealed by Over
1370 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**,
1371 649–662.e20 (2019).
1372 13. Vangay, P. *et al.* US Immigration Westernizes the Human Gut Microbiome. *Cell* **175**, 962–
1373 972.e10 (2018).
1374 14. Collinson, M. A. *et al.* Migration and the epidemiological transition: insights from the
1375 Agincourt sub-district of northeast South Africa. *Glob. Health Action* **7**, 23514 (2014).
1376 15. Griffiths, J. A. & Mazmanian, S. K. Emerging evidence linking the gut microbiome to
1377 neurologic disorders. *Genome Medicine* vol. 10 (2018).
1378 16. Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457**, 480–

1379 484 (2009).
1380 17. Helmink, B. A., Wadud Khan, M. A., Hermann, A., Gopalakrishnan, V. & Wargo, J. A. The
1381 microbiome, cancer, and cancer therapy. *Nature Medicine* vol. 25 377–388 (2019).
1382 18. Hagan, T. *et al.* Antibiotics-Driven Gut Microbiome Perturbation Alters Immunity to
1383 Vaccines in Humans. *Cell* **178**, 1313–1328.e13 (2019).
1384 19. Ciabattini, A., Olivieri, R., Lazzeri, E. & Medaglini, D. Role of the Microbiota in the
1385 Modulation of Vaccine Immune Responses. *Front. Microbiol.* **10**, 1305 (2019).
1386 20. Ou, J. *et al.* Diet, microbiota, and microbial metabolites in colon cancer risk in rural
1387 Africans and African Americans. *Am. J. Clin. Nutr.* **98**, 111–120 (2013).
1388 21. de la Cuesta-Zuluaga, J. *et al.* Gut microbiota is associated with obesity and
1389 cardiometabolic disease in a population in the midst of Westernization. *Sci. Rep.* **8**, 11356
1390 (2018).
1391 22. Jha, A. R. *et al.* Gut microbiome transition across a lifestyle gradient in Himalaya. *PLoS*
1392 *Biol.* **16**, e2005396 (2018).
1393 23. Lim, S. S. *et al.* A comparative risk assessment of burden of disease and injury attributable
1394 to 67 risk factors and risk factor clusters in 21 regions, 1990-2010: a systematic analysis
1395 for the Global Burden of Disease Study 2010. *Lancet* **380**, 2224–2260 (2012).
1396 24. Campbell, T. P. *et al.* The microbiome and resistome of chimpanzees, gorillas, and
1397 humans across host lifestyle and geography. *ISME J.* **14**, 1584–1599 (2020).
1398 25. Lokmer, A. *et al.* Use of shotgun metagenomics for the identification of protozoa in the gut
1399 microbiota of healthy individuals from worldwide populations with various industrialization
1400 levels. *PLoS One* **14**, e0211139 (2019).
1401 26. Tett, A. *et al.* The Prevotella copri Complex Comprises Four Distinct Clades
1402 Underrepresented in Westernized Populations. *Cell Host Microbe* **26**, 666–679.e7 (2019).
1403 27. Rocafort, M. *et al.* Evolution of the gut microbiome following acute HIV-1 infection.
1404 *Microbiome* vol. 7 (2019).
1405 28. Jacobson, D. K. *et al.* Analysis of global human gut metagenomes shows that metabolic
1406 resilience potential for short-chain fatty acid production is strongly influenced by lifestyle.
1407 *Sci. Rep.* **11**, 1724 (2021).
1408 29. Yinda, C. K. *et al.* Gut Virome Analysis of Cameroonians Reveals High Diversity of Enteric
1409 Viruses, Including Potential Interspecies Transmitted Viruses. *mSphere* **4**, (2019).
1410 30. Oduaran, O. H. *et al.* Gut Microbiome Profiling of a Rural and Urban South African Cohort
1411 Reveals Biomarkers of a Population in Lifestyle Transition. *Biorxiv* (2020)
1412 doi:10.1101/2020.02.27.964023.

1413 31. Santosa, A. & Byass, P. Diverse Empirical Evidence on Epidemiological Transition in Low-
1414 and Middle-Income Countries: Population-Based Findings from INDEPTH Network Data.
1415 *PLoS One* **11**, e0155753 (2016).
1416 32. Kabudula, C. W. *et al.* Progression of the epidemiological transition in a rural South African
1417 setting: findings from population surveillance in Agincourt, 1993--2013. *BMC Public Health*
1418 **17**, 424 (2017).
1419 33. Ajayi, I. O. *et al.* Urban-rural and geographic differences in overweight and obesity in four
1420 sub-Saharan African adult populations: a multi-country cross-sectional study. *BMC Public*
1421 *Health* **16**, 1126 (2016).
1422 34. NCD Risk Factor Collaboration (NCD-RisC) – Africa Working Group. Trends in obesity and
1423 diabetes across Africa from 1980 to 2014: an analysis of pooled population-based studies.
1424 *Int. J. Epidemiol.* **46**, 1421–1432 (2017).
1425 35. Sonnenburg, E. D. & Sonnenburg, J. L. The ancestral and industrialized gut microbiota and
1426 implications for human health. *Nat. Rev. Microbiol.* **17**, 383–390 (2019).
1427 36. Statistics South Africa. Census 2011 Statistical Release. (2012).
1428 37. Houle, B., Clark, S. J., Gómez-Olivé, F. X., Kahn, K. & Tollman, S. M. The unfolding
1429 counter-transition in rural South Africa: mortality and cause of death, 1994-2009. *PLoS*
1430 *One* **9**, e100420 (2014).
1431 38. Bawah, A. *et al.* The Evolving Demographic and Health Transition in Four Low- and
1432 Middle-Income Countries: Evidence from Four Sites in the INDEPTH Network of
1433 Longitudinal Health and Demographic Surveillance Systems. *PLoS One* **11**, e0157281
1434 (2016).
1435 39. Ginsburg, C. *et al.* Migration and Settlement Change in South Africa: Triangulating Census
1436 2011 with Longitudinal Data from the Agincourt Health and Demographic Surveillance
1437 System in the Rural North-east. *South. Afr. J. Demogr.* **17**, 133–198 (2016).
1438 40. De Filippo, C. *et al.* Impact of diet in shaping gut microbiota revealed by a comparative
1439 study in children from Europe and rural Africa. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 14691–
1440 14696 (2010).
1441 41. Gorvitovskaia, A., Holmes, S. P. & Huse, S. M. Interpreting Prevotella and Bacteroides as
1442 biomarkers of diet and lifestyle. *Microbiome* **4**, 15 (2016).
1443 42. Maier, L. & Typas, A. Systematically investigating the impact of medication on the gut
1444 microbiome. *Curr. Opin. Microbiol.* **39**, 128–135 (2017).
1445 43. Vich Vila, A. *et al.* Impact of commonly used drugs on the composition and metabolic
1446 function of the gut microbiota. *Nat. Commun.* **11**, 362 (2020).

1447 44. Forslund, K. *et al.* Disentangling type 2 diabetes and metformin treatment signatures in the
1448 human gut microbiota. *Nature* **528**, 262–266 (2015).
1449 45. Bolourian, A. & Mojtahedi, Z. Streptomyces, shared microbiome member of soil and gut,
1450 as 'old friends' against colon cancer. *FEMS Microbiology Ecology* vol. 94 (2018).
1451 46. Soo, R. M. *et al.* An expanded genomic representation of the phylum cyanobacteria.
1452 *Genome Biol. Evol.* **6**, 1031–1045 (2014).
1453 47. Di Rienzi, S. C. *et al.* The human gut and groundwater harbor non-photosynthetic bacteria
1454 belonging to a new candidate phylum sibling to Cyanobacteria. *Elife* **2**, e01102 (2013).
1455 48. Guerin, E. *et al.* Biology and Taxonomy of crAss-like Bacteriophages, the Most Abundant
1456 Virus in the Human Gut. *Cell Host Microbe* (2018) doi:10.1016/j.chom.2018.10.002.
1457 49. Edwards, R. A. *et al.* Global phylogeography and ancient evolution of the widespread
1458 human gut virus crAssphage. *Nat Microbiol* **4**, 1727–1736 (2019).
1459 50. de la Cuesta-Zuluaga, J., Ley, R. E. & Youngblut, N. D. Struo: a pipeline for building
1460 custom databases for common metagenome profilers. *Bioinformatics* **36**, 2314–2315
1461 (2020).
1462 51. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny
1463 substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
1464 52. Beghini, F. *et al.* Integrating taxonomic, functional, and strain-level profiling of diverse
1465 microbial communities with bioBakery 3. *Elife* **10**, (2021).
1466 53. Pokusaeva, K., Fitzgerald, G. F. & van Sinderen, D. Carbohydrate metabolism in
1467 Bifidobacteria. *Genes & Nutrition* vol. 6 285–306 (2011).
1468 54. Oliphant, K. & Allen-Vercoe, E. Macronutrient metabolism by the human gut microbiome:
1469 major fermentation by-products and their impact on host health. *Microbiome* **7**, 91 (2019).
1470 55. Santoru, M. L. *et al.* Cross sectional evaluation of the gut-microbiome metabolome axis in
1471 an Italian cohort of IBD patients. *Scientific Reports* vol. 7 (2017).
1472 56. Ramsay, M. *et al.* H3Africa AWI-Gen Collaborative Centre: a resource to study the
1473 interplay between genomic and environmental risk factors for cardiometabolic diseases in
1474 four sub-Saharan African countries. *Glob Health Epidemiol Genom* **1**, e20 (2016).
1475 57. Bäckhed, F. *et al.* Dynamics and Stabilization of the Human Gut Microbiome during the
1476 First Year of Life. *Cell Host Microbe* **17**, 690–703 (2015).
1477 58. Sonnenburg, J. & Sonnenburg, E. A Microbiota Assimilation. *Cell Metab.* **28**, 675–677
1478 (2018).
1479 59. Nayfach, S., Shi, Z. J., Seshadri, R., Pollard, K. S. & Kyrpides, N. C. New insights from
1480 uncultivated genomes of the global human gut microbiome. *Nature* **568**, 505–510 (2019).

1481 60. Brown, C. T. & Irber, L. sourmash: a library for MinHash sketching of DNA. *JOSS* **1**, 27
1482 (2016).
1483 61. Koslicki, D. & Falush, D. MetaPalette: a -mer Painting Approach for Metagenomic
1484 Taxonomic Profiling and Quantification of Novel Strain Variation. *mSystems* **1**, (2016).
1485 62. Martínez, I. *et al.* The gut microbiota of rural papua new guineans: composition, diversity
1486 patterns, and ecological processes. *Cell Rep.* **11**, 527–538 (2015).
1487 63. Chen, L.-X., Anantharaman, K., Shaiber, A., Eren, A. M. & Banfield, J. F. Accurate and
1488 complete genomes from metagenomes. *Genome Res.* **30**, 315–333 (2020).
1489 64. Almeida, A. *et al.* A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499–
1490 504 (2019).
1491 65. Moss, E. L., Maghini, D. G. & Bhatt, A. S. Complete, closed bacterial genomes from
1492 microbiomes using nanopore sequencing. *Nat. Biotechnol.* (2020) doi:10.1038/s41587-
1493 020-0422-6.
1494 66. Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a
1495 metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**,
1496 725–731 (2017).
1497 67. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut
1498 microbiome. *Nat. Biotechnol.* **39**, 105–114 (2020).
1499 68. Han, C. *et al.* Complete genome sequence of Treponema succinifaciens type strain (6091).
1500 *Stand. Genomic Sci.* **4**, 361–370 (2011).
1501 69. Angelakis, E. *et al.* Treponema species enrich the gut microbiota of traditional rural
1502 populations but are absent from urban individuals. *New Microbes New Infect* **27**, 14–21
1503 (2019).
1504 70. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes
1505 substantially expands the tree of life. *Nat Microbiol* **2**, 1533–1542 (2017).
1506 71. Scher, J. U. *et al.* Expansion of intestinal Prevotella copri correlates with enhanced
1507 susceptibility to arthritis. *Elife* **2**, e01202 (2013).
1508 72. Stewart, R. D. *et al.* Compendium of 4,941 rumen metagenome-assembled genomes for
1509 rumen microbiome biology and enzyme discovery. *Nat. Biotechnol.* **37**, 953–961 (2019).
1510 73. Sato, M. P. *et al.* Comparison of the sequencing bias of currently available library
1511 preparation kits for Illumina sequencing of bacterial genomes and metagenomes. *DNA*
1512 *Res.* **26**, 391–398 (2019).
1513 74. Brito, I. L. *et al.* Transmission of human-associated microbiota along family and social
1514 networks. *Nat Microbiol* **4**, 964–971 (2019).

75. Rothschild, D. *et al.* Environment dominates over host genetics in shaping human gut microbiota. *Nature* **555**, 210–215 (2018).
76. Choudhury, A. *et al.* High-depth African genomes inform human migration and health. *Nature* **586**, 741–748 (2020).
77. Parks, D. H. *et al.* A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.* **38**, 1079–1086 (2020).
78. Tyler, A. D. *et al.* Evaluation of Oxford Nanopore's MinION Sequencing Device for Microbial Whole Genome Sequencing Applications. *Sci. Rep.* **8**, 10931 (2018).
79. Trönnberg, L., Hawksworth, D., Hansen, A., Archer, C. & Stenström, T. A. 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 (2010).
80. Leung, J. M., Graham, A. L. & Knowles, S. C. L. Parasite-Microbiota Interactions With the Vertebrate Gut: Synthesis Through an Ecological Lens. *Front. Microbiol.* **9**, 843 (2018).
81. Richter, L., Norris, S., Pettifor, J., Yach, D. & Cameron, N. Cohort Profile: Mandela's children: the 1990 Birth to Twenty study in South Africa. *Int. J. Epidemiol.* **36**, 504–511 (2007).
82. Kabudula, C. W. *et al.* 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 (2017).

Methods References

1. Ramsay, M. *et al.* H3Africa AWI-Gen Collaborative Centre: a resource to study the interplay between genomic and environmental risk factors for cardiometabolic diseases in four sub-Saharan African countries. *Glob Health Epidemiol Genom* **1**, e20 (2016).
2. Brewster, R. *et al.* Surveying Gut Microbiome Research in Africans: Toward Improved Diversity and Representation. *Trends Microbiol.* (2019) doi:10.1016/j.tim.2019.05.006.
3. Krueger, F. Trim Galore! http://www.bioinformatics.babraham.ac.uk/projects/trim_galore.
4. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* (2009).
5. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).

1546 6. de la Cuesta-Zuluaga, J., Ley, R. E. & Youngblut, N. D. Struo: a pipeline for building
1547 custom databases for common metagenome profilers. *Bioinformatics* **36**, 2314–2315
1548 (2020).
1549 7. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: estimating species
1550 abundance in metagenomics data. *PeerJ Comput. Sci.* **3**, e104 (2017).
1551 8. Beghini, F. *et al.* Integrating taxonomic, functional, and strain-level profiling of diverse
1552 microbial communities with bioBakery 3. *Elife* **10**, (2021).
1553 9. Brown, C. T. & Irber, L. sourmash: a library for MinHash sketching of DNA. *JOSS* **1**, 27
1554 (2016).
1555 10. Crusoe, M. R. *et al.* The khmer software package: enabling efficient nucleotide sequence
1556 analysis. *F1000Res.* **4**, 900 (2015).
1557 11. Kaminski, J. *et al.* High-Specificity Targeted Functional Profiling in Microbial Communities
1558 with ShortBRED. *PLoS Comput. Biol.* **11**, e1004557 (2015).
1559 12. McArthur, A. G. *et al.* The comprehensive antibiotic resistance database. *Antimicrob.*
1560 *Agents Chemother.* **57**, 3348–3357 (2013).
1561 13. Mallick, H. *et al.* Multivariable Association Discovery in Population-scale Meta-omics
1562 Studies. *bioRxiv* 2021.01.20.427420 (2021) doi:10.1101/2021.01.20.427420.
1563 14. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to
1564 single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
1565 15. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately
1566 reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165
1567 (2015).
1568 16. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat.*
1569 *Methods* **11**, 1144–1146 (2014).
1570 17. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to
1571 recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2016).
1572 18. Sieber, C. M. K. *et al.* Recovery of genomes from metagenomes via a dereplication,
1573 aggregation and scoring strategy. *Nat Microbiol* **3**, 836–843 (2018).
1574 19. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for
1575 genome assemblies. *Bioinformatics* vol. 29 1072–1075 (2013).
1576 20. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM:
1577 assessing the quality of microbial genomes recovered from isolates, single cells, and
1578 metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
1579 21. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069

1580 (2014).
1581 22. Laslett, D. & Canback, B. ARAGORN, a program to detect tRNA genes and tmRNA genes
1582 in nucleotide sequences. *Nucleic Acids Res.* **32**, 11–16 (2004).
1583 23. Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a
1584 metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**,
1585 725–731 (2017).
1586 24. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate
1587 genomic comparisons that enables improved genome recovery from metagenomes
1588 through de-replication. *ISME J.* **11**, 2864–2868 (2017).
1589 25. Moss, E. L., Maghini, D. G. & Bhatt, A. S. Complete, closed bacterial genomes from
1590 microbiomes using nanopore sequencing. *Nat. Biotechnol.* (2020) doi:10.1038/s41587-
1591 020-0422-6.
1592 26. Lin, Y. *et al.* Assembly of long error-prone reads using de Bruijn graphs. *Proc. Natl. Acad.*
1593 *Sci. U. S. A.* **113**, E8396–E8405 (2016).
1594 27. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection
1595 and genome assembly improvement. *PLoS One* **9**, e112963 (2014).
1596 28. Hunt, M. *et al.* Circlator: automated circularization of genome assemblies using long
1597 sequencing reads. *Genome Biol.* **16**, 294 (2015).
1598 29. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly
1599 from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
1600 30. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–
1601 3100 (2018).
1602 31. Lu, J. & Salzberg, S. L. SkewIT: The Skew Index Test for large-scale GC Skew analysis of
1603 bacterial genomes. *PLoS Comput. Biol.* **16**, e1008439 (2020).
1604 32. Carver, T., Thomson, N., Bleasby, A., Berriman, M. & Parkhill, J. DNAPlotter: circular and
1605 linear interactive genome visualization. *Bioinformatics* **25**, 119–120 (2009).
1606 33. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify
1607 genomes with the Genome Taxonomy Database. *Bioinformatics* (2019)
1608 doi:10.1093/bioinformatics/btz848.
1609 34. Kieft, K., Zhou, Z. & Anantharaman, K. VIBRANT: automated recovery, annotation and
1610 curation of microbial viruses, and evaluation of viral community function from genomic
1611 sequences. *Microbiome* **8**, 90 (2020).
1612 35. Gibson, M. K., Forsberg, K. J. & Dantas, G. Improved annotation of antibiotic resistance
1613 determinants reveals microbial resistomes cluster by ecology. *ISME J.* **9**, 207–216 (2015).

36. Asnicar, F. *et al.* Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. *Nat. Commun.* **11**, 2500 (2020).
37. Cole, J. R. *et al.* Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* **42**, D633–42 (2014).
38. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
39. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
40. R Core Team. R: A Language and Environment for Statistical Computing. (2019).
41. Venables, W. N. & Ripley, B. D. Modern Applied Statistics with S. (2002).
42. Ahlmann-Eltze, C. ggsignif: Significance Brackets for 'ggplot2'. (2019).
43. Kassambara, A. ggpubr: 'ggplot2' Based Publication Ready Plots. (2020).
44. Oksanen, J. *et al.* vegan: Community Ecology Package. (2019).
45. Paulson, J. N., Stine, O. C., Bravo, H. C. & Pop, M. Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods* **10**, 1200–1202 (2013).
46. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
47. Wilke, C. O. cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'. (2019).
48. Gentleman, R., Carey, V., Huber, W. & Hahne, F. genefilter: genefilter: methods for filtering genes from high-throughput experiments. (2019).
49. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. (2016).
50. Slowikowski, K. ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'. (2020).
51. Warnes, G. R., Bolker, B. & Lumley, T. gtools: Various R Programming Tools. (2020).
52. Gonçalves da Silva, A. harrietr: Wrangle Phylogenetic Distance Matrices and Other Utilities. (2017).
53. Wickham, H. Reshaping Data with the reshape Package. *J. Stat. Softw.* **21**, 1–20 (2007).
54. Wickham, H. *et al.* Welcome to the tidyverse. *Journal of Open Source Software* vol. 4 1686 (2019).