bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

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

[1]Department of Genetics, Stanford University, Stanford, CA, USA

[2]Sydney Brenner Institute for Molecular Bioscience, University of the Witwatersrand, Johannesburg, South Africa

[3]School of Medicine, Stanford University, Stanford, CA, USA

[4]Division of Human Genetics, School of Pathology, Faculty of Health Sciences, National Health Laboratory Service & University of the Witwatersrand, Johannesburg, South Africa

[5]SAMRC Developmental Pathways for Health Research Unit, Department of Paediatrics, University of the Witwatersrand, Johannesburg, South Africa

[6]School of Human Development and Health, University of Southampton, UK

[7]MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

[8]INDEPTH Network, East Legon, Accra, Ghana

[9]School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg, South Africa

[10]Department of Medicine (Hematology), Stanford University, Stanford, CA, USA

*Co-corresponding authors: asbhatt@stanford.edu, Scott.Hazelhurst@wits.ac.za

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

To date, human gut microbiome research in adults has focused on western populations, with an additional minority of studies examining nonwestern populations including agriculturalist or hunter-gatherer societies. Yet, the majority of the world's population resides along a continuum between these two extremes. South Africa is a country in rapid lifestyle and epidemiologic transition and is therefore a unique setting to explore uncharacterized features of the human gut microbiome. Here, we present the first study evaluating gut microbiome composition in transitional South African populations using both short- and long-read sequencing data, which enables both taxonomic profiling as well as generation of contiguous metagenomic assemblies. In this cross-sectional study, we sequenced and analyzed single stool samples from 168 adult females living either in rural Bushbuckridge municipality (n=117) or urban Soweto (n=51) in South Africa. We find that on the microbial community level, the gut microbiomes of South African individuals are consistent with an intermediate state between those of western and other previously studied African populations. Furthermore, we demonstrate that current reference collections are incomplete for nonwestern microbiomes and as a result, patterns of within-cohort beta diversity are reversed compared to the ground truth of reference-agnostic sequence comparison. To improve reference databases for South African populations, we generated several additional complete and contiguous draft genomes via assembly of short and long reads. We identify previously undescribed taxa, such as novel *Treponema*, *Lentisphaera* and *Succinatimonas* species, and report the first complete *Treponema succinifaciens* genome assembled from the gut of a human. Taken together, our results suggest that the transitional lifestyle and epidemiological conditions in South Africa contribute to gut microbiota compositions that are intermediate between western and non-western guts and that additional microbial genomic diversity remains to be described in these populations.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

Comprehensive characterization of the healthy human gut microbiota is essential to contextualize studies of the microbiome in disease. To date, substantial resources have been invested in describing the microbiome of individuals living in the global 'west' (United States, northern and western Europe), including efforts by large consortia such as the Human Microbiome Project (Human Microbiome Project Consortium, 2012) and metaHIT (Qin et al., 2010). Though these have yielded valuable descriptions of human gut microbial ecology, it is not clear to what extent these results are generalizable to non-western populations. A relatively smaller number of studies have characterized the gut microbiome composition of non-western individuals (Brewster et al., 2019; Gupta et al., 2017), including communities practicing traditional lifestyles in Venezuela and Malawi (Yatsunenko et al., 2012), hunter-gatherer communities in Tanzania (Fragiadakis et al., 2018; Rampelli et al., 2015; Schnorr et al., 2014; Smits et al., 2017), non-industrialized populations in Tanzania and Botswana (Hansen et al., 2019), and agriculturalists in remote Madagascar (Pasolli et al., 2019). However, many of the world's communities have lifestyles between the extremes of an urbanized lifestyle and traditional practices. Many communities are undergoing a transition of diet and lifestyle practice, characterized by increased access to processed foods, diets rich in animal fats and simple carbohydrates, and more sedentary lifestyles. This has corresponded with an epidemiological transition in which the burden of disease is shifting from predominantly infectious diseases to include increasing incidence of noncommunicable diseases like obesity and diabetes (Collinson et al., 2014). The microbiome has been implicated in various noncommunicable diseases (Griffiths and Mazmanian, 2018; Helmink et al., 2019; Turnbaugh et al., 2009) and may mediate the efficacy of medical interventions including vaccines (Ciabattini et al., 2019; Hagan et al., 2019), but we cannot evaluate the generalizability of these findings without establishing baseline microbiome characteristics of communities that are transitioning between traditional and western lifestyles. These understudied populations offer a unique opportunity to examine the relationship between lifestyle (including diet), disease, and gut microbiome composition, and to discover novel microbial genomic content.

South Africa is a prime example of a country undergoing rapid lifestyle and epidemiological transition. With the exception of the HIV/AIDS epidemic in the mid-1990s to the mid-2000s, South Africa has experienced a steadily decreasing mortality from infectious disease and an increase in noncommunicable disease over the past three decades (Kabudula et

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

al., 2017a; Santosa and Byass, 2016). Concomitantly, increasingly sedentary lifestyles and changes in dietary habits, including access to calorie-dense processed foods, contribute to a higher prevalence of obesity in many regions of South Africa (Kabudula et al., 2017a), a trend which disproportionately affects women (Ajayi et al., 2016; NCD Risk Factor Collaboration (NCD-RisC) – Africa Working Group, 2017).

Previous studies have begun to probe the relationship between diet, lifestyle, and microbiome composition in transitional communities. Rural South Africans eating their typical fiber-rich diet were found to have increased relative abundance of *Prevotella* and butyrate-producing bacteria and lower *Bacteroides* compared to African Americans (Gupta et al., 2017; Ou et al., 2013), while a survey of the gut microbiome composition in a "westernizing" community of adults living in rural Colombia found that these individuals harbor taxa classically associated with nonwestern communities (*Prevotella*, *Treponema*) as well as western communities (*Bacteroides*, *Bifidobacterium*) (de la Cuesta-Zuluaga et al., 2018). Further, a study of four Himalayan populations in various stages of a transition from hunting and gathering to agriculturalist lifestyles revealed a corresponding shift in gut microbiome composition (Jha et al., 2018). However, substantial gaps remain in our description of the microbiome in transitional communities. Many studies of the human gut microbiome in transitioning communities have focused on children, whose disease risk profile and gut microbiome composition can vary considerably from adults (Lim et al., 2012; Yatsunenko et al., 2012). Additionally, studies of the gut microbiome in South Africa and other transitional populations to date have used 16S rRNA sequencing, which is limited to the description of relative abundances of bacteria and archaea that have been previously described and deposited into reference collections. Further, 16S rRNA sequencing lacks genomic resolution to resolve strains of the same species that have the same 16S rRNA sequence but differ substantially in gene content or genome structure.

To address this knowledge gap, we designed and performed the first research study applying short- and long-read DNA sequencing to study the gut microbiomes of South African individuals for whom 16S rRNA sequence data has recently been reported (Oduaran et al., 2020). In this work, we describe microbial community-scale similarities between urban and rural communities in South Africa, as well as distinct hallmark taxa that distinguish each community. Additionally, we place South Africans in context with microbiome data from other populations globally, revealing the transitional nature of gut microbiome composition in the South African cohorts. We demonstrate that metagenomic assembly of short reads yields novel strain and species draft genomes. Finally, we apply Oxford Nanopore long-read sequencing to samples

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

from the rural cohort and generate complete and near-complete genomes. These include genomes of species that are exclusive to, or more prevalent in, traditional populations, including *Treponema* and *Prevotella* species. As long-read sequencing enables more uniform coverage of AT-rich regions compared to short-read sequencing with transposase-based library preparation, we also generate complete metagenome-assembled AT-rich genomes from less well-described gut microbes including species in the phylum *Melainabacteria*, the class *Mollicutes*, and the genus *Mycoplasma*. Taken together, the results herein offer a more detailed description of gut microbiome composition in understudied transitioning populations, and present complete and contiguous reference genomes that will enable further studies of gut microbiota in nonwestern populations.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Results

We enrolled 190 women aged between 40-60, living in rural villages in the Bushbuckridge Municipality (24.8°E, 31.3°S, n=132) and urban Soweto, Johannesburg (26.2°S, 27.9°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. Only samples from HIV-negative individuals were analyzed further (n=117 Bushbuckridge, n=51 Soweto). 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 35.8 million (M) raw reads were generated per sample (range 11.4 M - 100 M), and a median of 12 M reads (range 3.2 M - 29.3 M) resulted after pre-processing including de-duplication, trimming, and human read removal (Table S1).

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 sequencing of the same samples (Oduaran et al., 2020), we find that *Prevotella*, *Faecalibacterium*, and *Bacteroides* are the most abundant genera in most individuals across both study sites (Figure 1A, Figure S1, Table S2; species-level classifications in Table S3). 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, Paraprevotellaceae, and Spirochaetaceae) (Fragiadakis et al., 2018) such as *Prevotella, Treponema*, and *Succinatimonas*, which have been demonstrated to be higher in relative abundance in communities practicing traditional lifestyles compared to westerners (Fragiadakis et al., 2018; Sonnenburg and Sonnenburg, 2019) (Figure 1B, Table S2). Mean relative abundance of each VANISH genus is higher in Bushbuckridge than Soweto, though the difference is not statistically significant for *Prevotella*, *Paraprevotella*, or *Alkalispirochaeta* (Figure 1B, Wilcoxon rank-sum test). Within the Bushbuckridge cohort, we observe a bimodal distribution of the genera *Succinatimonas*, *Succinivibrio*, and *Treponema* (Figure S2A). While we do not identify any participant metadata that associate with this distribution, we observe that VANISH taxa are weakly correlated in metagenomes from both Bushbuckridge and Soweto (Figure S2B-C). We identify the bacteriophage crAssphage and related crAss-like phages (Guerin et al., 2018), which have recently been described as prevalent constituents of the gut microbiome

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

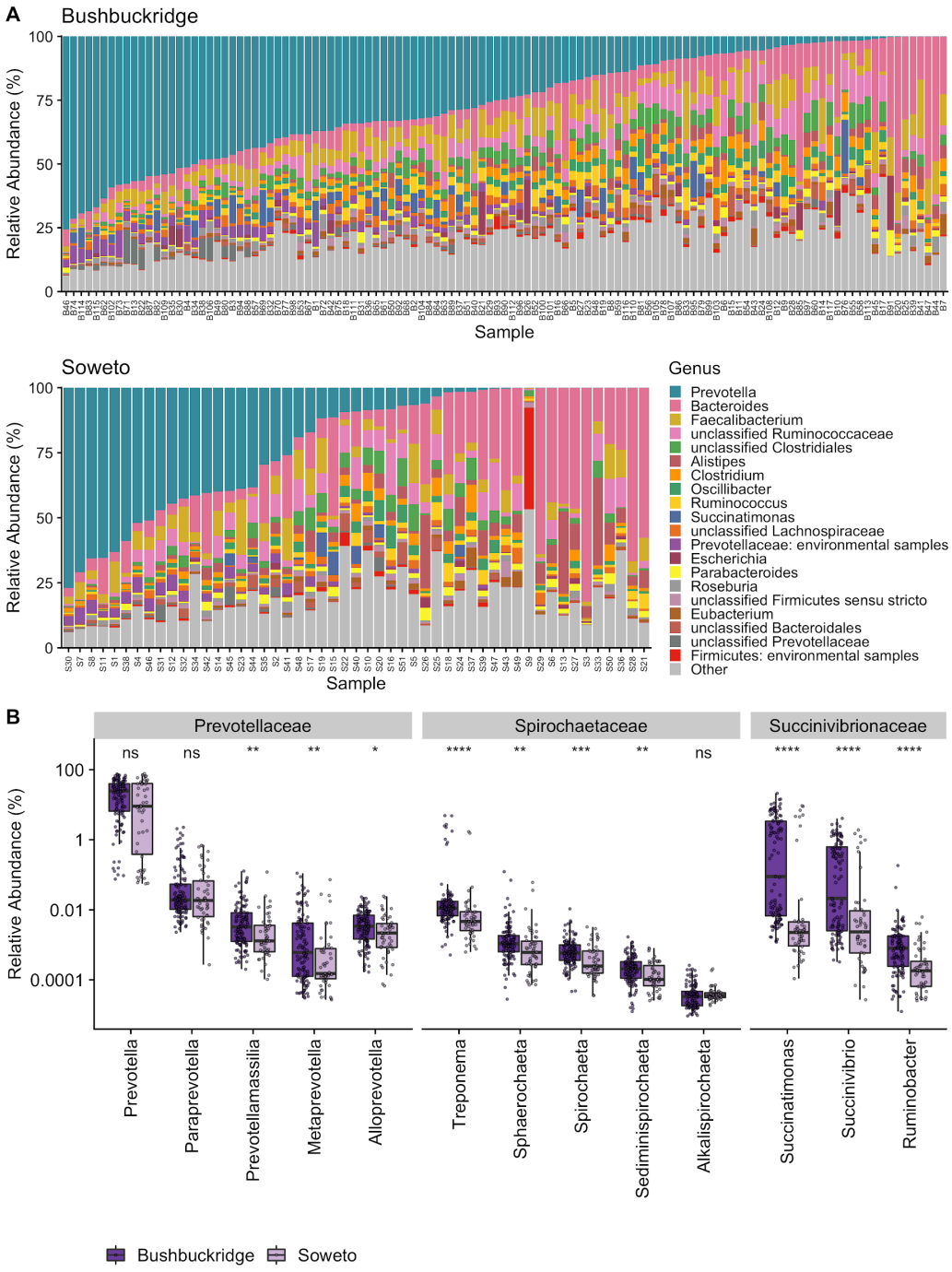


Figure 1. Taxonomic composition of South African study participants

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

(A) Top 20 genera by relative abundance for samples from participants in Bushbuckridge and Soweto, sorted by decreasing *Prevotella* abundance. *Prevotella*, *Faecalibacterium*, and *Bacteroides* are the most prevalent genera across both study sites.

(B) Relative abundance of VANISH genera by study site, grouped by family. A pseudocount of 1 read was added to each sample prior to relative abundance normalization in order to plot on a log scale, as the abundance of some genera in some samples is zero. Relative abundance of most VANISH genera is higher on average in participants from Bushbuckridge than Soweto (Wilcoxon rank-sum test, significance values denoted as follows: (*) $p < 0.05$, (**) $p < 0.01$, (***) $p < 0.001$, (****) $p < 0.0001$, (ns) not significant.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

globally (Edwards et al., 2019), in 116 of 168 individuals (69%) (using 650 sequence reads or roughly 1X coverage of the 97 kb genome as a threshold for crAss-like phage presence).

Rural and urban microbiomes cluster distinctly in MDS: We hypothesized that lifestyle differences of those residing in rural Bushbuckridge versus urban Soweto might be associated with demonstrable differences in gut microbiome composition. Bushbuckridge and Soweto differ markedly in their population density (53 and 6,357 persons per km^2 respectively as of the 2011 census) as well as in lifestyle variables including the prevalence of flush toilets (6.8 vs 91.6% of dwellings) and piped water (11.9 vs 55% of dwellings) (additional site demographic information in Table S4) (Statistics South Africa, 2012). Soweto is highly urbanized and has been so for several generations, while Bushbuckridge is classified as a rural community, although it is undergoing rapid epidemiological transition. Bushbuckridge also sees circular rural/urban migrancy typified by some members of a rural community working and living for extended periods in urban areas, while keeping their permanent rural home: although our participants all live in Bushbuckridge, this migrancy in the community helps make the boundary between rural and urban lifestyles more fluid. Comparing the two study populations at the community level, we find that samples from the two sites have distinct centroids (PERMANOVA $p < 0.001$, $R^2 = 0.037$) but overlap (Figure 2A), though we note that the dispersion of the Soweto samples is greater than that of the Bushbuckridge samples (PERMDISP2 $p < 0.001$). Across the study population we observe a gradient of Bacteroides and Prevotella relative abundance (Figure S3). This is likely a result of differences in diet across the study population at both sites, as Bacteroides and Prevotella have been proposed as biomarkers of diet and lifestyle (De Filippo et al., 2010; Gorvitovskaia et al., 2016; Yatsunenko et al., 2012).

Rural and urban microbiomes differ in Shannon diversity and species composition: The gut microbiome Shannon diversity of individuals living traditional lifestyles is higher than those living western lifestyles. In keeping with this general trend, we find that alpha diversity (Shannon) is significantly higher in individuals living in rural Bushbuckridge than urban Soweto (Figure 2B; Wilcoxon rank-sum test, $p < 0.01$). Using DESeq2 to identify microbial genera that are differentially abundant across study sites, we find that genera including *Bacteroides*, *Bifidobacterium,* and *Staphylococcus* are more abundant in individuals living in Soweto (Figure 2C, Table S5, species shown in Figure S4). Interestingly, we find microbial genera enriched in gut microbiomes of individuals living in Bushbuckridge which are common to both the environment and the gut, including *Streptomyces* and *Pseudomonas* (Table S5). Typically a soil-associated organism, *Streptomyces* encode a variety of biosynthetic gene clusters and can

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

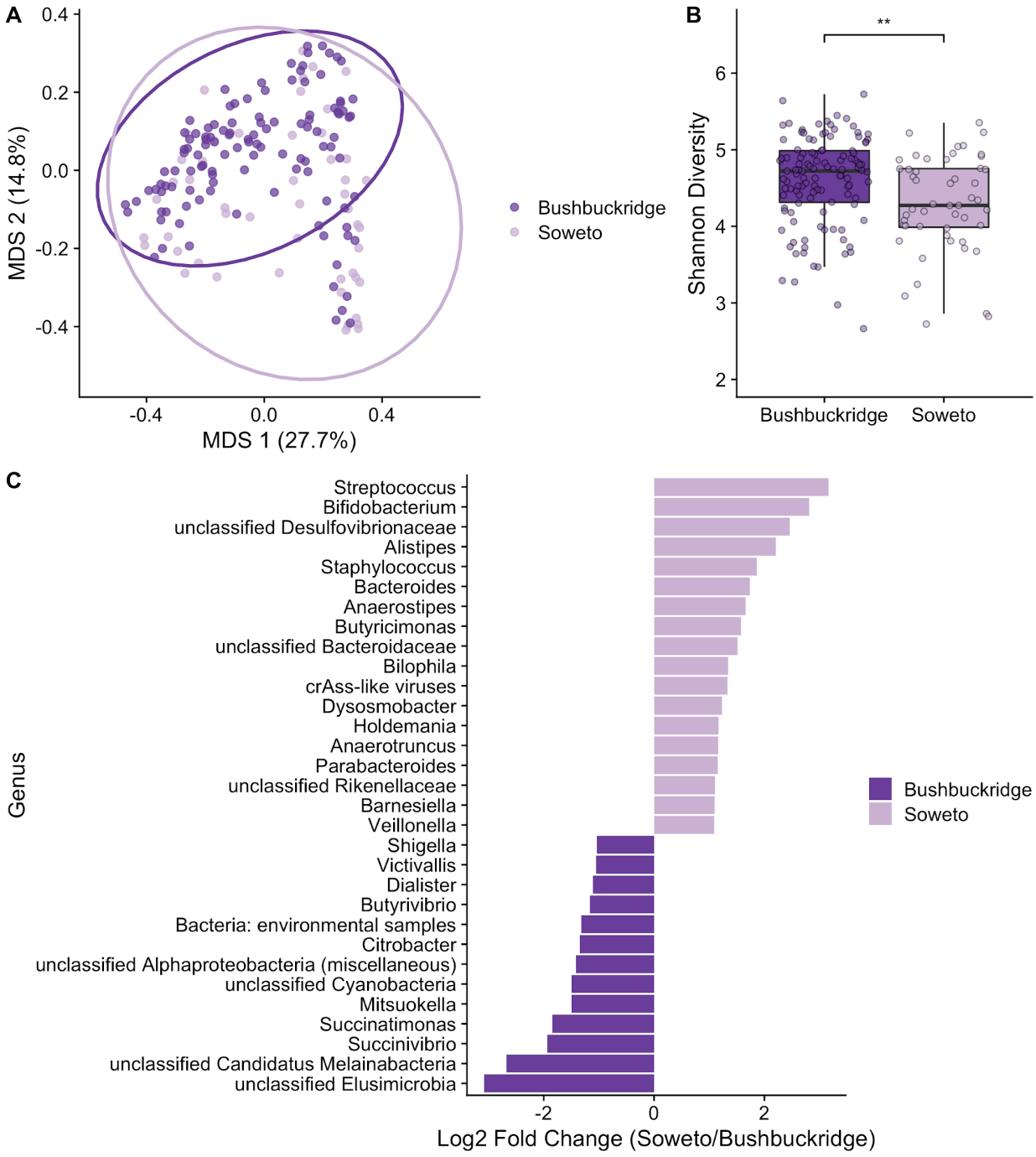


Figure 2. Comparison of Bushbuckridge and Soweto microbiomes

(A) Multidimensional scaling of pairwise Bray-Curtis distance between samples. Samples from Soweto are more dispersed than samples from Bushbuckridge ($p < 0.001$).

(B) Shannon diversity calculated on species-level taxonomic classifications for each sample. Samples from Bushbuckridge are higher in alpha diversity than samples from Soweto (Wilcoxon rank-sum test, $p < 0.001$).

(C) DESeq2 identifies microbial genera that are differentially abundant in rural Bushbuckridge compared to the urban Soweto cohort. Features with log2 fold change greater than one are plotted (full results in Table S5).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

produce numerous immunomodulatory and anti-inflammatory compounds such as rapamycin and tacrolimus, and it has been suggested that decreased exposure to *Streptomyces* is associated with increased incidence of inflammatory disease and colon cancer in western populations (Bolourian and Mojtahedi, 2018). In addition, we find enrichment in various other genera in Bushbuckridge including *Succinatimonas*, a relatively poorly-described bacterial genus with only one type species, that has been previously associated with nonwestern microbiomes (Pasolli et al., 2019), and Candidatus Melainabacter, a phylum closely related to Cyanobacteria that in limited studies have been described to inhabit the human gut (Di Rienzi et al., 2013; Soo et al., 2014)

No strong signals of interaction between human DNA variation and microbiome content detected: We have a very small sample size to assess interaction between human genetic variation and microbiome population. However, as our study is one of the relatively few with both human and microbiome DNA characterized, we performed association tests between key microbiome genera abundance levels and the human DNA. After correcting for multiple testing there were only a few SNPs with borderline statistically significant association with genera abundance levels (Table S6). They occur in genomic regions with no obvious impact on the gut microbiome (see *Methods/Supplementary Information*). Additionally, we do not observe that samples cluster by self-reported ethnicity of the participant (Figure S5).

South African gut microbiomes are intermediate in composition between western and nonwestern populations: To place the microbiome composition of South African individuals in global context with metagenomes from healthy adults living in other parts of the world, we compared publicly available data from four cohorts (Figure 3A, Table S7) comprising adult individuals living in the United States (Human Microbiome Project Consortium, 2012), northern Europe (Sweden) (Bäckhed et al., 2015), rural Madagascar (Pasolli et al., 2019), as well as the Hadza hunter-gatherers of Tanzania (Rampelli et al., 2015). We note the caveat that these samples were collected at different times using different approaches, and that there is variation in DNA extraction, sequencing library preparation and sequencing, all of which may contribute to variation between studies. Recognizing this limitation, we observe that South African samples cluster between western and nonwestern populations in MDS (Figure 3B) as expected, and that the first axis of MDS correlates well with geography and lifestyle (Figure 3C). Additionally, Bacteroidaceae, Prevotellaceae, and Succinivibrionaceae are strongly correlated with the first axis of MDS (Spearman's rho > 0.8): Bacteroidaceae decreases with MDS 1 while Prevotellaceae and Succinivibrionaceae increase (Figure 3B). These observations suggest that

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

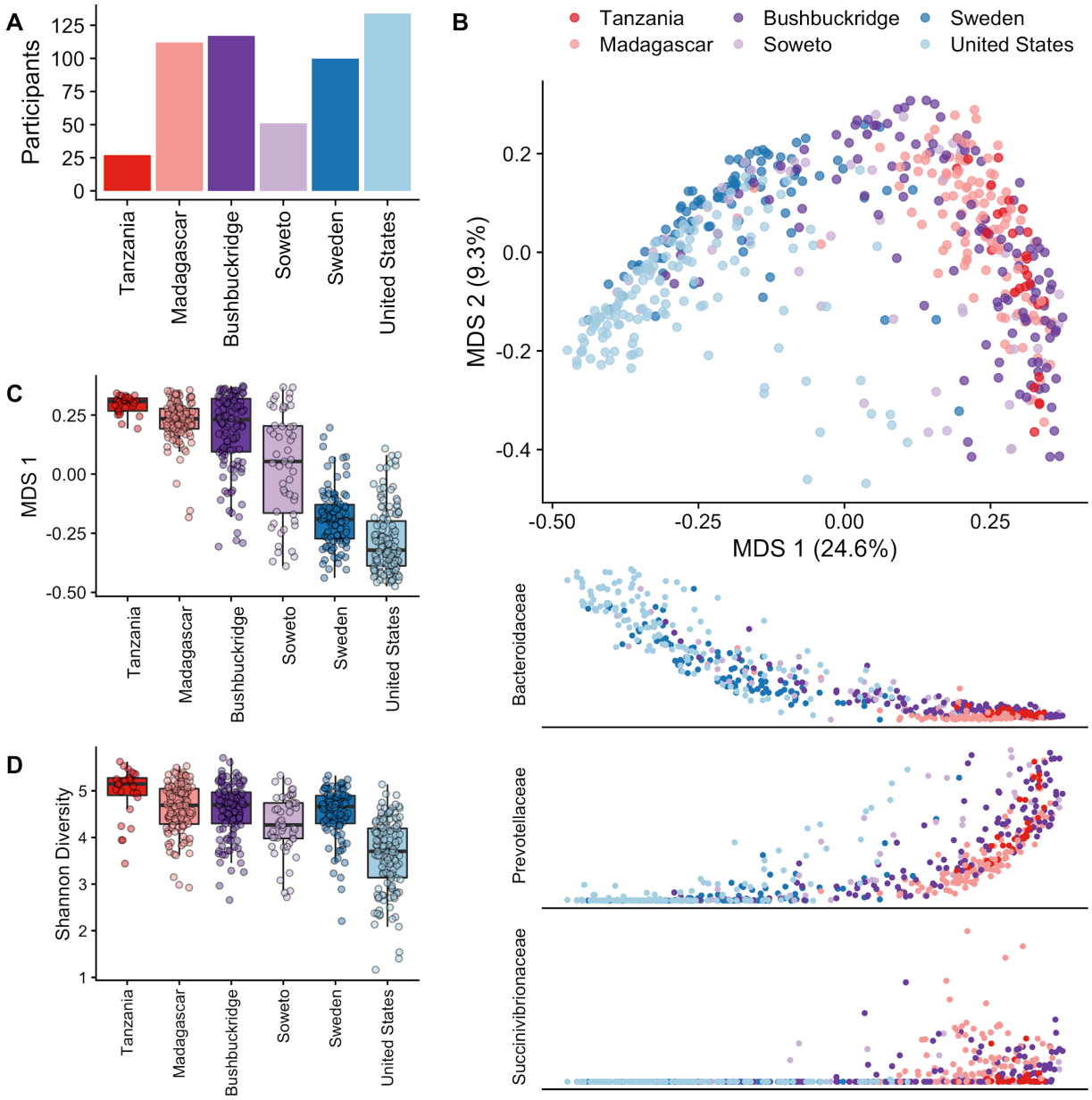


Figure 3. Community-level comparison of global microbiomes

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

(A) Number of participants per cohort.

(B) Multidimensional scaling of pairwise Bray-Curtis distance between samples from six datasets of healthy adult shotgun microbiome sequencing data. Western populations (Sweden, United States) cluster away from African populations practicing a traditional lifestyle (Madagascar, Tanzania) while transitional South African microbiomes overlap with both western and nonwestern populations. Shown below are scatterplots of relative abundance of the top three taxa most correlated with MDS 1 (Spearman's rho, Bacteroidaceae -0.89, Prevotellaceae 0.87, Succinivibrionaceae 0.82), against MDS 1 on the x axis.

(C) Boxplot of the first axis of MDS (MDS 1) which correlates with geography and lifestyle.

(D) Shannon diversity across cohorts. Shannon diversity was calculated from data rarefied to the number of sequence reads of the lowest sample. Shannon diversity is highest in nonwestern cohorts (Tanzania, Madagascar), lowest in western populations (United States, Sweden) and intermediate in South Africans (Bushbuckridge, Soweto).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the transitional lifestyle of South African individuals is reflected in their gut microbiome composition. We observe a corresponding pattern of decreasing relative abundance of VANISH taxa across lifestyle and geography (Figure S6). In addition to differences in species composition, we see evidence that populations vary on the strain level: we annotated the pangenome (collection of open reading frames of all strains of a clade) of the most abundant species across the combined datasets and observe that for some species, pangenomes from Western individuals tend to cluster together, while pangenomes from South Africans, Tanzanians, and Madagascans cluster separately (Figure 4)

Decreased sequence classifiability in nonwestern populations: Given previous observations that gut microbiome alpha diversity is higher in individuals practicing traditional lifestyles (Gupta et al., 2017; Smits et al., 2017; Sonnenburg and Sonnenburg, 2018) and that immigration from a nonwestern nation to the United States is associated with a decrease in gut microbial alpha diversity (Vangay et al., 2018), we hypothesized that alpha diversity would be higher in nonwestern populations including South Africans. We observe that Shannon diversity of the Tanzanian hunter-gatherer cohort is uniformly higher than all other populations (Figure 3D; $p < 0.01$ for all pairwise comparisons; FDR-adjusted Wilcoxon rank sum test) and that alpha diversity is lower in individuals living in the United States compared to all other cohorts (Figure 3D; $p < 0.0001$ for all pairwise comparisons; FDR-adjusted Wilcoxon rank sum test). Surprisingly, we observe comparable Shannon diversity between Madagascar, Bushbuckridge, and Sweden (ns, Wilcoxon rank sum test). However, this could be a result of incomplete representation of diverse microbes in existing reference collections.

Classification of metagenomic sequences from nonwestern gut microbiomes with existing reference collections is known to be limited (Nayfach et al., 2019; Pasolli et al., 2019), and we observe decreased sequence classifiability in nonwestern populations (Figure S7A). Therefore, we sought orthogonal validation of our observation that South African microbiomes represent a transitional state between traditional and western microbiomes and employed a reference-independent method to evaluate the nucleotide composition of sequence data from each metagenome. We used the sourmash workflow (Brown and Irber, 2016) to compare nucleotide *k*-mer composition of sequencing reads in each sample and ordinated based on Jaccard distance. Using a *k*-mer length of 31 (high k-mer similarity at k=31 correlates with species-level similarity (Koslicki and Falush, 2016)), we observe clustering reminiscent of the species ordination plot shown in Fig. 3, further supporting the hypothesis that South African

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

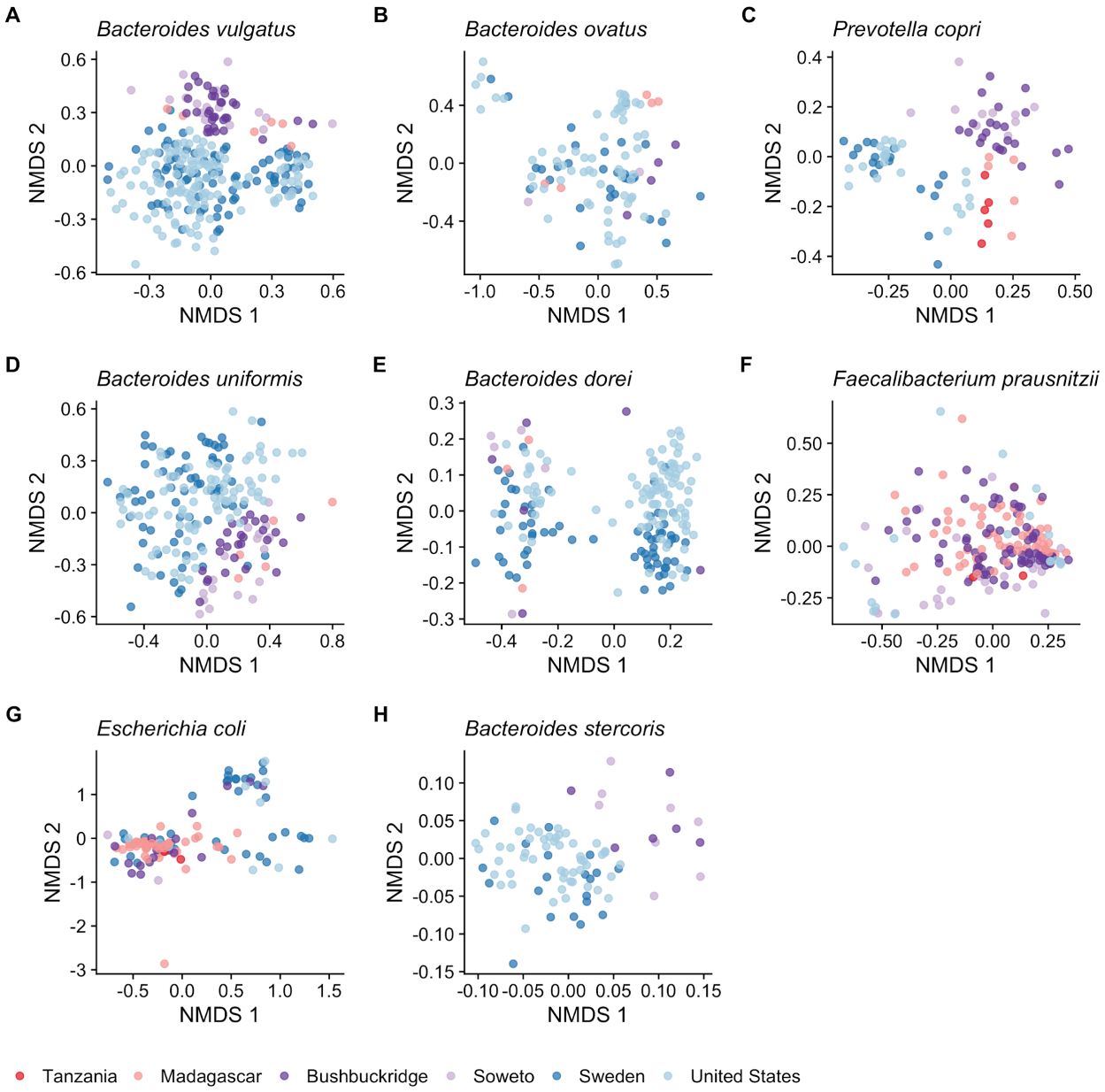


Figure 4. Pangenomes of shared species differ between western and nonwestern microbiomes

Pangenomes of the eight species with the highest average relative abundance across cohorts were determined for each sample using PanPhlAn, and the resulting matrix of gene presence/absence was used to calculate Jaccard distance and ordinate via nonmetric multidimensional scaling. Species were selected on the basis of high relative abundance across multiple cohorts. Pangenomes of samples from nonwestern populations cluster away from western populations for some species (A, C, D, G, H).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

microbiomes are transitional (Figure S7C, comparisons at additional lengths of *k* shown in Figure S7B,D). Additionally, we observe greater beta diversity within nonwestern populations relative to westerners (Figure S7C). This observation is critically important to our understanding of beta diversity in the gut microbiome in western and nonwestern communities, as previous studies have observed a pattern of higher alpha diversity but lower beta-diversity in nonwestern populations compared to western populations (Martínez et al., 2015; Schnorr et al., 2014). We observe that this holds true when comparing the distributions of pairwise Bray-Curtis distances (based on species classification) between populations but not when comparing nucleotide *k*-mers (Figure S7D-E). This suggests that the generalization of an inverse relationship between alpha and beta diversity is perhaps overly simplistic and potentially an artifact of limitations in classification databases.

Improving reference collections via metagenomic assembly: Classification of metagenomic sequencing reads can be improved by assembling sequencing data into metagenomic contigs and grouping these contigs into draft genomes (binning), yielding metagenome-assembled genomes (MAGs). The majority of publications to date have focused on creating MAGs from short-read sequencing data (Almeida et al., 2019; Nayfach et al., 2019; Pasolli et al., 2019), but generation of high-quality MAGs from long-read data from stool samples has been recently reported (Moss et al., 2020). To better characterize the genomes present in our samples, we assembled and binned shotgun sequencing reads from South African samples into MAGs (Figure S8). We generated 3312 MAGs (43 high-quality, 1510 medium-quality, and 1944 low-quality) (Bowers et al., 2017) from 168 metagenomic samples, which yielded a set of 1192 non-redundant medium-quality or better representative strain genomes when filtered for completeness greater than 50%, and contamination less than 10% and de-replicated at 99% average nucleotide identity (ANI). This collection of dereplicated genomes includes VANISH taxa including *Prevotella*, *Treponema*, and *Sphaerochaeta* species (Figure S9, Table S8).

Interestingly, many MAGs within this set represent organisms that are uncommon in Western microbiomes or not easily culturable, including organisms from the genera *Treponema* and *Vibrio*. As short-read MAGs are typically fragmented and exclude mobile genetic elements, we explored methods to create more contiguous genomes, with a goal of trying to better understand these understudied taxa. We performed long-read sequencing on three samples from participants in Bushbuckridge with an Oxford Nanopore MinION sequencer (taxonomic composition of the three samples shown in Figure S10). Samples were chosen for nanopore

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

sequencing on the basis of molecular weight distribution and total mass of DNA (see Methods). One flow cell per sample generated an average 19.71 Gbp of sequencing with a read N50 of 8,275 bp after basecalling. From our three samples, we generated 741 nanopore MAGs (nMAGs), which yielded 35 non-redundant genomes when filtered for completeness greater than 50% and contamination less than 10%, and dereplicated at 99% ANI (Figure 5, Table 1, Figure S8, Table S9). All of the dereplicated nMAGs contained at least one full length 16S sequence, and 28 nMAGs had contig N50s of greater than 1 Mbp (Table S9).

We compared assembly statistics between all MAGs and nMAGs, and found that while nMAGs were typically evaluated as less complete by CheckM, the contiguity of nanopore medium- and high-quality MAGs was an order of magnitude higher (mean nMAG N50 of 260.5 kb compared to mean N50 of medium- and high-quality MAGs of 15.1 kb) at comparable levels of average coverage (Figure S8, Figure S11). We expect that CheckM under-calculates the completeness of nanopore MAGs due to the homopolymer errors common in nanopore sequencing, which result in frameshift errors when annotating genomes. Indeed, we observe that nanopore MAGs of comparably high assembly size and low contamination as short-read MAGs are evaluated by CheckM as having lower completeness (Figure S11).

Novel genomes generated through nanopore sequencing: Comparisons between the full set of nMAGs and MAGs with publicly available genomes in GenBank reveals that many of the South African MAGs share less than 95% ANI to known species. When filtered for alignments covering at least 60% of the nMAG or MAG length, 107 nMAGs and 133 MAGs have less than 95% ANI to any genome in GenBank, while 93 nMAGs and 14 MAGs share less than 95% ANI to any genome in previously published MAG databases (Almeida et al., 2019; Nayfach et al., 2019; Pasolli et al., 2019) (Fig. 5A). Some of these organisms are likely absent from reference databases because they are uncommon in western populations or are difficult to culture. For example, we assembled two single contig, megabase-scale genomes from the genus *Treponema,* a clade that contains various commensal and pathogenic species and is uncommon in the gut microbiota of western individuals (Obregon-Tito et al., 2015; Schnorr et al., 2014). The first of these genomes is a single-contig *Treponema succinifaciens* genome. The type strain of *T. succinifaciens*, isolated from the swine gut (Han et al., 2011), is the only genome of this species currently available in public reference collections. Our *T. succinifaciens* genome is the first complete genome of this species from the gut of a human. We assembled a second *Treponema* sp. (Figure S12), which shares 92.1% ANI with *T. succinifaciens*. Additionally, we assembled a 5.08 Mbp genome for *Lentisphaerae sp*., which has been shown to be significantly enriched in

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

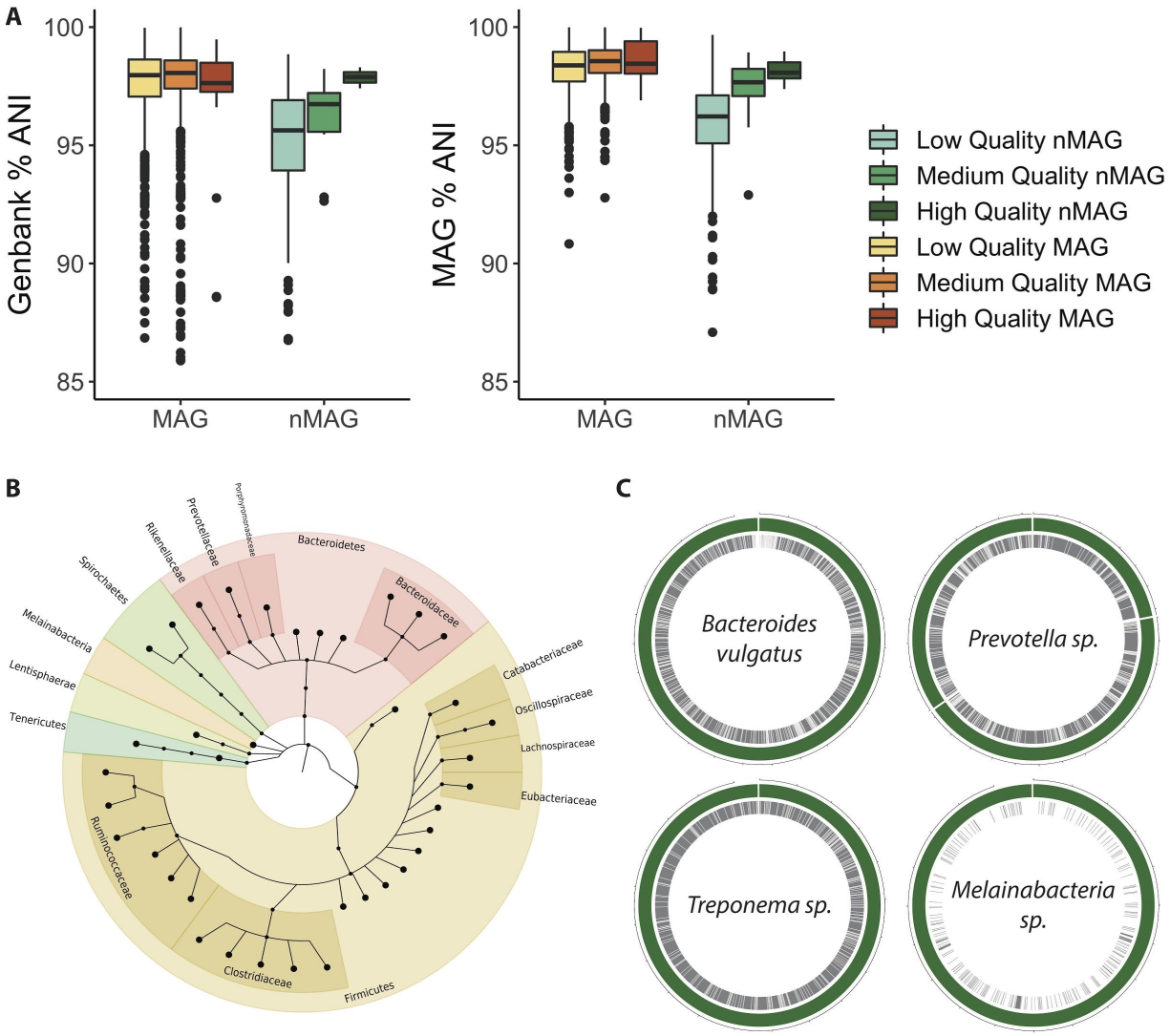


Figure 5. Complete and contiguous genomes of South African microbiota

(A) Percent ANI between South African MAGs and the highest identity GenBank genomes and previously published MAGs. Points are filtered to only show South African MAGs with alignments covering >60% of the assembled South African MAG genome.

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

(C) A selection of MAGs assembled from long-read sequencing (green) of three South African samples compared contigs assembled from corresponding short read data (grey). Breaks in circles represent different contigs.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

335 traditional populations (Angelakis et al., 2019). Our genome has 94% 16S rRNA identity with
336 *Victivallis vadensis*, suggesting a new species or genus of the family *Victivallaceae* and
337 representing the second closed genome for the phylum *Lentisphaerae*.
338 Other nMAGs represent organisms that are prevalent in western individuals but
339 challenging to assemble due to their genome structure. Despite the prevalence of *Bacteroides* in
340 western microbiomes, only three closed *B. vulgatus* genomes are available in RefSeq. We
341 assembled a single contig, 2.68 Mbp *Bacteroides vulgatus* genome that is 65.0% complete and
342 2.7% contaminated and contains at least 16 putative insertion sequences, which may contribute
343 to its lack of contiguous short-read assemblies. Similarly, we assembled a single contig genome
344 for *Catabacter sp.*, a member of the order *Clostridiales*; the most contiguous *Catabacter* genome
345 in GenBank is in five scaffolded contigs (Parks et al., 2017). The putative *Catabacter sp.* shares
346 85% ANI with the best match in GenBank, suggesting that it represents a new species within the
347 *Catabacter* genus or a new genus entirely. Additionally, we assembled a 3.6 Mbp genome for
348 *Prevotella sp.* (N50 = 1.87 Mbp), a highly variable genus that is prevalent in nonwestern
349 microbiomes and associated with a range of effects on host health (Scher et al., 2013). Notably,
350 the first closed genomes of *P. copri*, a common species of *Prevotella*, were only recently
351 assembled with nanopore sequencing of metagenomic samples; one from a human stool sample
352 and the other from cow rumen (Moss et al., 2020; Stewart et al., 2019). *P. copri* had previously
353 evaded closed assembly from short-read sequencing due to the dozens of repetitive insertion
354 sequences within its genome (Moss et al., 2020).
355 We observed that many long-read assembled genomes were evaluated as having low
356 completeness despite having contig N50 values greater than 1 Mbp. In investigating this
357 phenomenon, we discovered that many of these genomes had sparse or uneven short-read
358 coverage, leading to gaps in short-read polishing that would otherwise correct small frameshift
359 errors. To polish genomic regions that were not covered with short-reads, we performed long-
360 read polishing on assembled contigs from each sample, and re-binned polished contigs. Long-
361 read polishing improved the completeness of many organisms that are not commonly described
362 in the gut microbiota, due perhaps to their low relative abundance in the average human gut, or
363 to biases in shotgun sequencing library preparation that limit their detection (Figure S13, Figure
364 S14). For example, we generated a 2 Mbp genome that is best classified as a species of the
365 phylum Melainabacteria. Melainabacteria is a non-photosynthetic phylum closely related to
366 Cyanobacteria that has been previously described in the gut microbiome and is associated with
367 consuming a vegetarian diet (Di Rienzi et al., 2013). Melainabacteria have proven difficult to

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

isolate and culture, and the only complete, single-scaffold genome exists in RefSeq was assembled from shotgun sequencing of a human fecal sample (Di Rienzi et al., 2013). Interestingly, our Melainabacteria genome has a GC content of 30.9%, and along with assemblies of a *Mycoplasma sp.* (25.3% GC) and *Mollicutes sp. (*28.1% GC) (Figure S15), represent AT-rich organisms that can be underrepresented in shotgun sequencing data due to the inherent GC bias of transposon insertion and amplification-based sequencing approaches (Sato et al., 2019) (Figure S14). Altogether, these three genomes increased in completeness by an average of 28.5% with long-read polishing to reach an overall average of 70.88% complete. While these genomes meet the accepted standards to be considered medium-quality, it is possible that some or all of these highly contiguous, megabase scale assemblies are complete or near-complete yet underestimated by CheckM due to incomplete polishing.

Altogether, we find that *de novo* assembly approaches are capable of generating contiguous, high-quality assemblies for novel organisms, offering potential for investigation into the previously unclassified matter in the microbiomes of these nonwestern communities. In particular, nanopore sequencing was able to produce contiguous genomes for organisms that are difficult to assemble due to repeat structures (*Prevotella sp.*, *Bacteroides vulgatus*), as well as for organisms that exist on the extreme ends of the GC content spectrum (*Mollicutes sp.*, *Melainabacteria sp.*). We observe that long-reads are able to capture a broader range of taxa both at the read and assembly levels when compared to short-read assemblies, and that short- and long-read polishing approaches are able to yield medium-quality or greater draft genomes for these organisms. This illustrates the increased visibility that *de novo* assembly approaches lend to the study of the full array of organisms in the gut microbiome.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Discussion

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

We discover that gut microbiome composition differs demonstrably between the Bushbuckridge and Soweto cohorts, further highlighting the importance of studying diverse communities with differing lifestyle practices. Interestingly, even though gut microbiomes of individuals in Bushbuckridge and Soweto share many features and are more similar to each other than to other global cohorts studied, we do observe hallmark taxa associated with westernization are enriched in microbiomes in Soweto. These include *Bacteroides* and *Bifidobacterium*, which have been previously associated with urban communities (Gupta et al., 2017), consistent with Soweto's urban locale in the Johannesburg metropolitan area. Conversely, individuals in Bushbuckridge are enriched in VANISH taxa including *Succinatimonas*, which has been recently reported as associated with microbiomes from individuals practicing traditional lifestyles (Pasolli et al., 2019). Intriguingly, several VANISH taxa (*Succinatimonas*, *Succinivibrio*, *Treponema*) display bimodal distributions in the Bushbuckridge cohort. We hypothesize that the differences in distribution within and between cohorts is related to differences in lifestyle factors including dietary preferences rather than host genetics.

Despite the fact that host genetics explain relatively little of the variation in microbiome composition (Rothschild et al., 2018), we do observe a small number of taxa that associate with host genetics in this population. Future work is required for replication and to determine whether these organisms are interacting with the host and whether they are associated with host health.

Additionally, we demonstrate marked differences between South African cohorts and other previously studied populations living on the African continent and western countries. Broadly, we find that South African microbiomes reflect the transitional nature of their communities in that they overlap with western and nonwestern populations. We also find that species common to western and nonwestern populations vary widely in their gene complement. Tremendous human genetic diversity exists within Africa, and our work reveals that there is a great deal of as yet unexplored microbiome diversity as well. In fact, we find that microbiome

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

beta diversity within communities may be systematically underestimated by incomplete reference databases: taxa that are unique to individuals in nonwestern populations are not present in reference databases and therefore not included in beta diversity calculations. Though it has been reported that nonwestern and traditional populations tend to have higher alpha diversity but lower beta diversity compared to western populations, we show that this pattern is not upheld when reference-agnostic nucleotide-based comparisons are applied. By extension, we speculate that previous claims that beta-diversity inversely correlates with alpha diversity may have been fundamentally limited by study designs. Specifically, the disparity between comparing small, homogenous African populations with large, heterogenous western ones constitutes a significant statistical confounder. This prevents a valid assessment of beta-diversity between groups.

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

Additionally, we find that while our short-read draft genomes have closely related genomes in public reference databases, many nanopore draft genomes are more distant from their closest related genomes, suggesting that nanopore sequencing can capture a broad array of taxa. We produced genomes for organisms that exist on the extremes of the GC content spectrum, such as *Mycoplasma sp.*, *Mollicutes sp.*, and *Melainabacteria sp*. We find that these organisms are sparsely covered by short-read sequencing, illustrating the increased range of non-

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

amplification based sequencing approaches, such as nanopore sequencing. Interestingly, these assemblies are evaluated as only medium-quality by CheckM despite having low measurements of contamination, as well as genome lengths and gene counts comparable to reference genomes from the same phylogenetic clade. We hypothesize that sparse short-read coverage leads to incomplete polishing and therefore retention of small frameshift errors, which are a known limitation of nanopore sequencing (Tyler et al., 2018). Further evaluation of 16S or long-read sequencing of traditional and western populations can identify whether these organisms are specific to certain lifestyles, or more prevalent but poorly detected with shotgun sequencing.

While we find that the gut microbiome composition of the two South African cohorts described herein reflects their lifestyle transition, we acknowledge that these cohorts are not necessarily representative of all transitional communities in South Africa or other parts of the world which differ in lifestyle, diet, and resource access. Hence, further work remains to describe the gut microbiota in detail of other such understudied populations. This includes a detailed characterization of parasites present in microbiome sequence data, an analysis which we did not undertake in this study but which would be of great interest. These organisms have been detected in the majority of household toilets in nearby KwaZulu-Natal province (Trönnberg et al., 2010), and may interact with and influence microbiota composition (Leung et al., 2018).

Our study has several limitations. Although the publicly available sequence data from other global cohorts were generated with similar methodology to our study, it is possible that batch effects exist between datasets generated in different laboratories that may explain some percentage of the global variation we observe. Additionally, while nanopore sequencing is able to broaden our range of investigation, we illustrate that our ability to produce well-polished genomes at GC content extremes is limited. This may affect our ability to accurately call gene lengths and structures, although iterative long-read polishing improves our confidence in these assemblies. Future investigation of these communities using less biased, higher coverage short-read approaches or more accurate long-read sequencing approaches, such as PacBio circular consensus sequencing, may improve assembly qualities. Additionally, long-read sequencing of samples from a wider range of populations can identify whether the genomes identified herein are limited to traditional and transitional populations, or more widespread. Further, future improvements in error rate of long-read sequencing may obviate the need for short-read polishing altogether.

Taken together, our results emphasize the importance of generating sequence data from diverse transitional populations to contextualize studies of health and disease in these

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

individuals. To do so with maximum sensitivity and precision, reference genomes must be generated to classify sequencing reads from these metagenomes. Herein, we demonstrate the discrepancies in microbiome sequence classifiability across global populations and highlight the need for more comprehensive reference collections. Recent efforts have made tremendous progress in improving the ability to classify microbiome data through creating new genomes via metagenomic assembly (Almeida et al., 2019; Nayfach et al., 2019; Pasolli et al., 2019), and here we demonstrate the application of short- and long-read metagenomic assembly techniques to create additional genome references. Our application of long-read sequencing technology to samples from South African individuals has demonstrated the ability to generate highly contiguous MAGs and shows immense potential to expand our reference collections and better describe microbiomes throughout diverse populations globally.

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Methods

Cohort selection

Stool samples were collected from women aged 40-60 years in Soweto, South Africa and Bushbuckridge, South Africa. Human subjects research approval was obtained (Stanford IRB 43069, University of the Witwatersrand Human Research Ethics Committee M160121, Mpumalanga Provincial Health Research Committee MP_2017RP22_851) and informed consent was obtained from participants for all samples collected. Stool samples were collected and preserved in OmniGene Gut OMR-200 collection kits (DNA Genotek). Samples were stored at room temperature in the collection kit for 4-8 weeks until aliquoting followed by long-term storage at -80°C. As the enrollment criteria for our study included previous participation in a larger human genomics project (Ramsay et al., 2016), we had access to self-reported ethnicity for each participant (BaPedi, Ndebele, Sotho, Tsonga, Tswana, Venda, Xhosa, Zulu, Other, or Unknown). Samples from participants who tested HIV-positive or who did not consent to an HIV test were not analyzed.

Metagenomic sequencing of stool samples

DNA was extracted from stool samples using the QIAamp PowerFecal DNA Kit (QIAGEN) according to the manufacturer's instructions except for the lysis step, in which samples were lysed using the TissueLyser LT (QIAGEN) (30 second oscillations/3 minutes at 30Hz). DNA concentration of all DNA samples was measured using Qubit Fluorometric Quantitation (DS DNA High-Sensitivity Kit, Life Technologies). DNA sequencing libraries were prepared using the Nextera XT DNA Library Prep Kit (Illumina). Final library concentration was measured using Qubit Fluorometric Quantitation and library size distributions were analyzed with the Bioanalyzer 2100 (Agilent). Libraries were multiplexed and 150 base pair paired-end reads were generated on the HiSeq 4000 platform (Illumina). Samples with greater than approximately 300 ng remaining mass and a peak fragment length of greater than 19,000 bp (with minimal mass under 4,000 bp) as determined by a TapeStation 2200 (Agilent Technologies, Santa Clara, CA) were selected for nanopore sequencing. Nanopore sequencing libraries were prepared using the 1D Genomic DNA by Ligation protocol (ONT, Oxford UK) following standard instructions. Each library was sequenced with a full FLO-MIN106D R9 Version Rev D flow cell on a MinION sequencer for at least 60 hours.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Computational methods

Preprocessing

Stool metagenomic sequencing reads were trimmed using TrimGalore v0.5.0 (Krueger), a wrapper for CutAdapt v1.18 (Martin, 2011), with a minimum quality score of 30 for trimming (--q 30) and minimum read length of 60 (--length 60). Trimmed reads were deduplicated to remove PCR and optical duplicates using seqtk rmdup v1.3-r106 with default parameters. Reads aligning to the human genome (hg19) were removed using bwa v0.7.17-r1188 (Li and Durbin, 2009). To assess the microbial composition of our short-read sequencing samples, we used the Kraken v2.0.8-beta taxonomic sequence classifier with default parameters (Wood and Salzberg, 2014) and a comprehensive custom reference database containing all bacterial and archaeal genomes in GenBank assembled to "complete genome," "chromosome," or "scaffold" quality as of January 2020. Bracken v2.0.0 was then used to re-estimate abundance at each taxonomic rank (Lu et al., 2017).

Additional data

Published data from additional populations were downloaded via the NCBI Sequence Read Archive (SRA) or European Nucleotide Archive (Table S7) and preprocessed and taxonomically classified as described above. For datasets containing longitudinal samples from the same individual, one unique sample per individual was chosen (the first sample from each individual was chosen from the United States Human Microbiome Project cohort).

K-mer sketches

K-mer sketches were computed using sourmash v2.0.0a11 (Brown and Irber, 2016). Low abundance *k*-mers were trimmed using the "trim-low-abund.py" script from the khmer package (Crusoe et al., 2015) with a k-mer abundance cutoff of 3 (-C 3) and trimming coverage of 18 (-Z 18). Signatures were computed for each sample using the command "sourmash compute" with a compression ratio of 1000 (--scaled 1000) and *k*-mer lengths of 21, 31, and 51 (-k 21,31,51). Signatures at each length of *k* were compared using "sourmash compare" with default parameters and the correct length of *k* specified with the -k flag.

Pangenome generation

PanPhlAn v1.2.2.5 with default settings was used to reconstruct species-specific pangenomes (Scholz et al., 2016).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Statistical analysis and plotting

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

Genome assembly, binning, and evaluation

Short-read metagenomic data were assembled with MEGAHIT v1.1.3 (Li et al., 2016) and binned into draft genomes as previously described (Bishara et al., 2018). Briefly, short reads were aligned to assembled contigs with BWA v0.7.17 (Li and Durbin, 2009) and contigs were subsequently binned into draft genomes with MetaBAT v2:2.13 (Kang et al., 2015). Bins were evaluated for size, contiguity, completeness, and contamination with QUAST v5.0.0 (Gurevich et al., 2013), checkM v1.0.13 (Parks et al., 2015), Prokka v1.13 (Seemann, 2014), Aragorn v1.2.38 (Laslett and Canback, 2004), and Barrnap v0.9 (https://github.com/tseemann/barrnap/). We referred to published guidelines to designate genome quality (Bowers et al., 2017). Individual contigs from all assemblies were assigned taxonomic classifications with Kraken v2.0.8 (Bowers et al., 2017; Wood and Salzberg, 2014). Genome sets were filtered for completeness greater than 50% and contamination less than 10% as evaluated by checkM, and dereplicated using dRep v2.5.4 (Olm et al., 2017) with ANI threshold to form secondary clusters (-sa) at 0.99 (strain-level) or 0.95 (species-level).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

ANI comparisons were performed between South African MAGs and genomes available in GenBank (Nov. 2019) and genomes published in three recent MAG publications (Almeida et al., 2019; Nayfach et al., 2019; Pasolli et al., 2019). The top fifty reference genome matches were found by first computing Mash distance between South African MAGs and MinHash sketches of the GenBank and MAG databases, followed by nucmer (Kurtz et al., 2004) comparisons between top hits to identify the best GenBank and MAG reference. Taxonomic trees were plotted with Graphlan v1.1.3 (Asnicar et al., 2015).

Data availability

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

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 DHPRU 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

aspects of the project. We thank the Stanford Research Computing Center and Ben Siranosian for their contributions to computational infrastructure and support.

Funding

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 R01 AI148623 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. The work was further supported by the South African National Research Foundation (CPRR160421162721) and a seed grant from the African Partnership for Disease Control. O.H.O. was partially supported by a Fogarty Global Health Equity Scholar award (TW009338). The AWI-Gen project is supported by the National Human Genome Research Institute (U54HG006938) as part of the H3A Consortium. ANW is supported by the Fogarty International Centre, National Institutes of Health under award number K43TW010698. This paper describes the views of the authors and does not necessarily represent the official views of the National Institutes of Health (USA)

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Main Tables

Table 1. Medium- and high-quality genomes assembled from nanopore sequencing

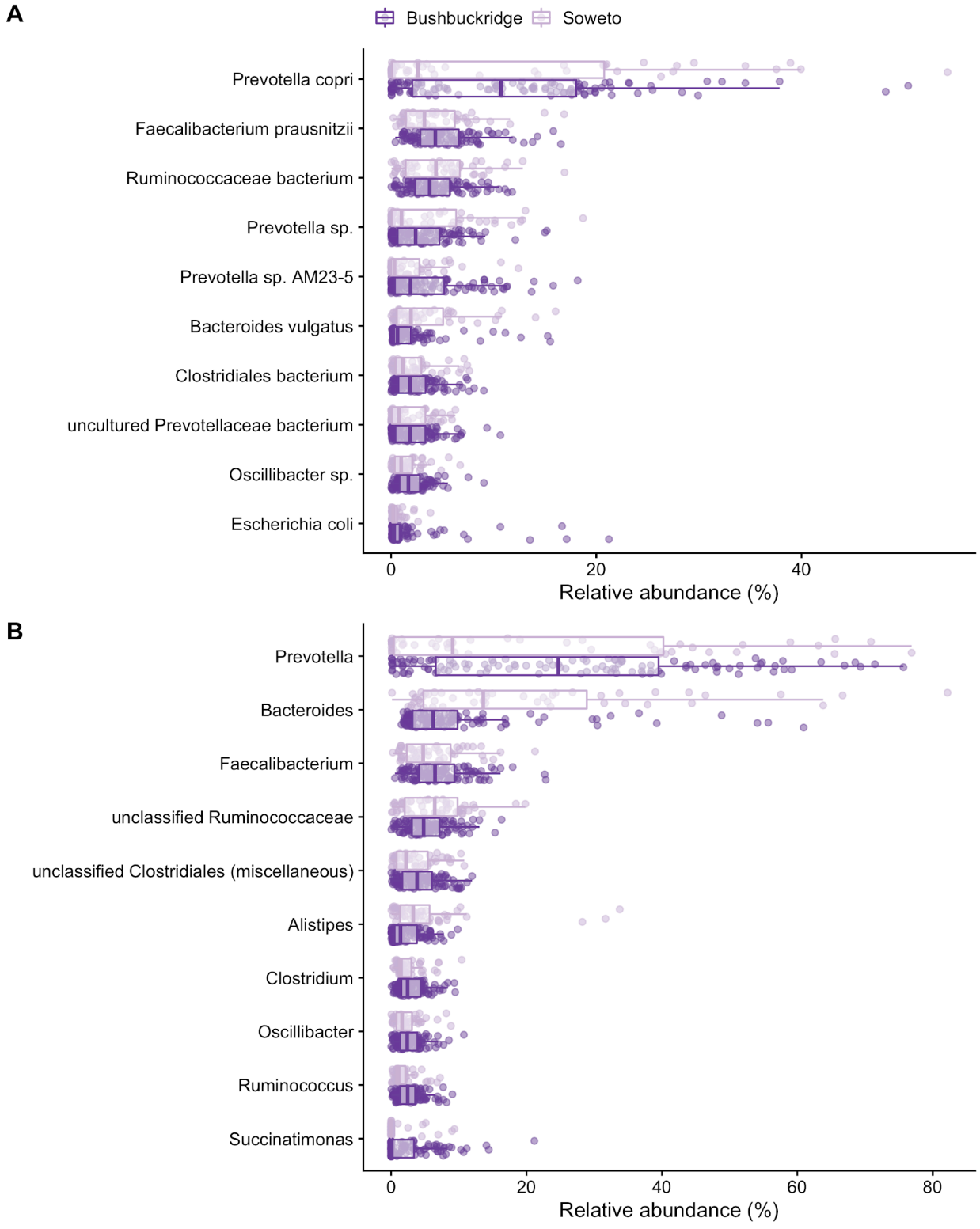
Classification	Size (Mb)	N50 (Mb)	GC	Completeness	Contamination	Quality	16S	23S	5S	tRNA	Polishing
Alistipes_putredinis	1.91	1.91	53.1	81.6	0.43	Medium	2	2	2	42	Short Read Only
Anaerotruncus_sp.	2.04	2.04	43.7	82.9	0	Medium	2	2	3	52	Short Read Only
Bacilli_bacterium	1.46	1.46	26.2	71.1	1.12	Medium	1	1	1	32	Short Read Only
Bacteroidales_bacterium	1.7	1.70	56.6	52.1	0.48	Medium	1	1	1	29	Short Read Only
Bacteroides_sp.	2.82	2.00	43.3	68.9	0.19	Medium	6	6	6	61	Short Read Only
Bacteroides_vulgatus	2.68	2.68	42.7	65.0	2.69	Medium	3	3	3	62	Short Read Only
Candidatus_Melainabacteria	2	2.00	30.9	71.4	1.23	Medium	1	1	1	36	Long and Short Read
Catabacter_sp.	1.65	1.65	46.4	83.4	2.68	Medium	1	1	1	48	Long and Short Read
Clostridiales_bacterium	2.03	0.60	57.9	57.3	0.7	Medium	4	4	5	39	Short Read Only
Clostridiales_bacterium	1.95	0.73	49.6	80.9	1.34	Medium	3	3	3	51	Short Read Only
Clostridiales_bacterium	2.24	0.58	48.7	68.0	1.51	Medium	2	2	2	41	Short Read Only
Clostridiales_bacterium	2.65	2.65	42.8	53.4	0.06	Medium	3	3	3	45	Short Read Only
Clostridiales_bacterium	1.32	0.79	45.2	62.5	1.21	Medium	1	2	2	30	Short Read Only
Clostridiales_bacterium	1.53	1.53	47.3	77.7	1.34	Medium	1	1	1	46	Short Read Only
Clostridiales_bacterium	1.61	1.61	46.9	72.1	0.7	Medium	1	1	1	46	Short Read Only
Clostridium_sp.	2.01	2.01	28.8	77.3	1.33	Medium	3	3	3	37	Short Read Only
Clostridium_sp.	1.53	1.53	25.2	63.3	0.37	Medium	1	1	1	33	Short Read Only
Clostridium_sp.	1.3	1.30	46.9	62.9	0.67	Medium	1	1	1	40	Short Read Only
Clostridium_sp.	1.14	1.14	29.1	66.9	0	Medium	1	1	1	34	Short Read Only
Eubacterium	2	0.63	44.5	62.4	0.73	Medium	2	2	2	38	Short Read Only
Lachnospiraceae_bacterium	3.38	1.94	43.6	88.2	1.55	Medium	4	4	0	53	Short Read Only
Lachnospiraceae_bacterium	3.81	2.83	43.6	79.0	1.75	Medium	4	4	0	53	Short Read Only
Lentisphaeria_bacterium	5.08	5.08	57.5	81.9	2.14	Medium	3	3	3	56	Long and Short Read
Mollicutes_bacterium	1.68	1.49	28.1	66.8	2.79	Medium	2	2	2	36	Long and Short Read
Mycoplasma_sp.	1.17	1.12	25.3	74.4	1.12	Medium	2	2	0	31	Long and Short Read
Oscillibacter_sp.	1.13	0.17	57.4	53.2	0	Medium	1	1	1	21	Short Read Only
Porphyromonadaceae_bacterium	2.97	2.97	47.4	83.5	0.13	Medium	5	5	5	59	Short Read Only
Prevotella_sp.	3.29	1.14	43.6	63.1	2	Medium	6	6	6	61	Long and Short Read
Ruminococcaceae_bacterium	1.95	0.80	38.4	74.2	1.34	Medium	4	4	5	58	Short Read Only
Ruminococcaceae_bacterium	1.78	1.78	58.3	51.6	0	Medium	3	3	4	42	Short Read Only
Treponema_sp.	2.06	2.06	41.6	81.5	0	Medium	3	3	3	44	Short Read Only
uncultured_Ruminococcus	1.59	1.34	44.0	66.3	1.01	Medium	2	2	1	32	Short Read Only
uncultured_Ruminococcus	2.08	2.08	46.9	66.4	0	Medium	5	5	5	53	Short Read Only
Bacteroidales_bacterium	2.79	2.79	49.8	94.1	0	High	4	4	4	57	Short Read Only

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Bacteroidales_bacterium	2.67	1.80	47.3	94.0	0.54	High	3	3	3	46	Short Read Only
Bacteroides_sp.	2	1.59	48.2	92.1	0.48	High	3	3	3	42	Short Read Only
Clostridium_sp.	2.44	2.23	52.5	91.9	0	High	3	3	3	55	Short Read Only
Ruminococcaceae_bacterium	2.27	2.27	51.4	90.4	0.34	High	3	3	3	51	Short Read Only
Treponema_succinifaciens	2.55	2.55	39.1	94.8	0.7	High	4	4	5	49	Short Read Only

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Figures



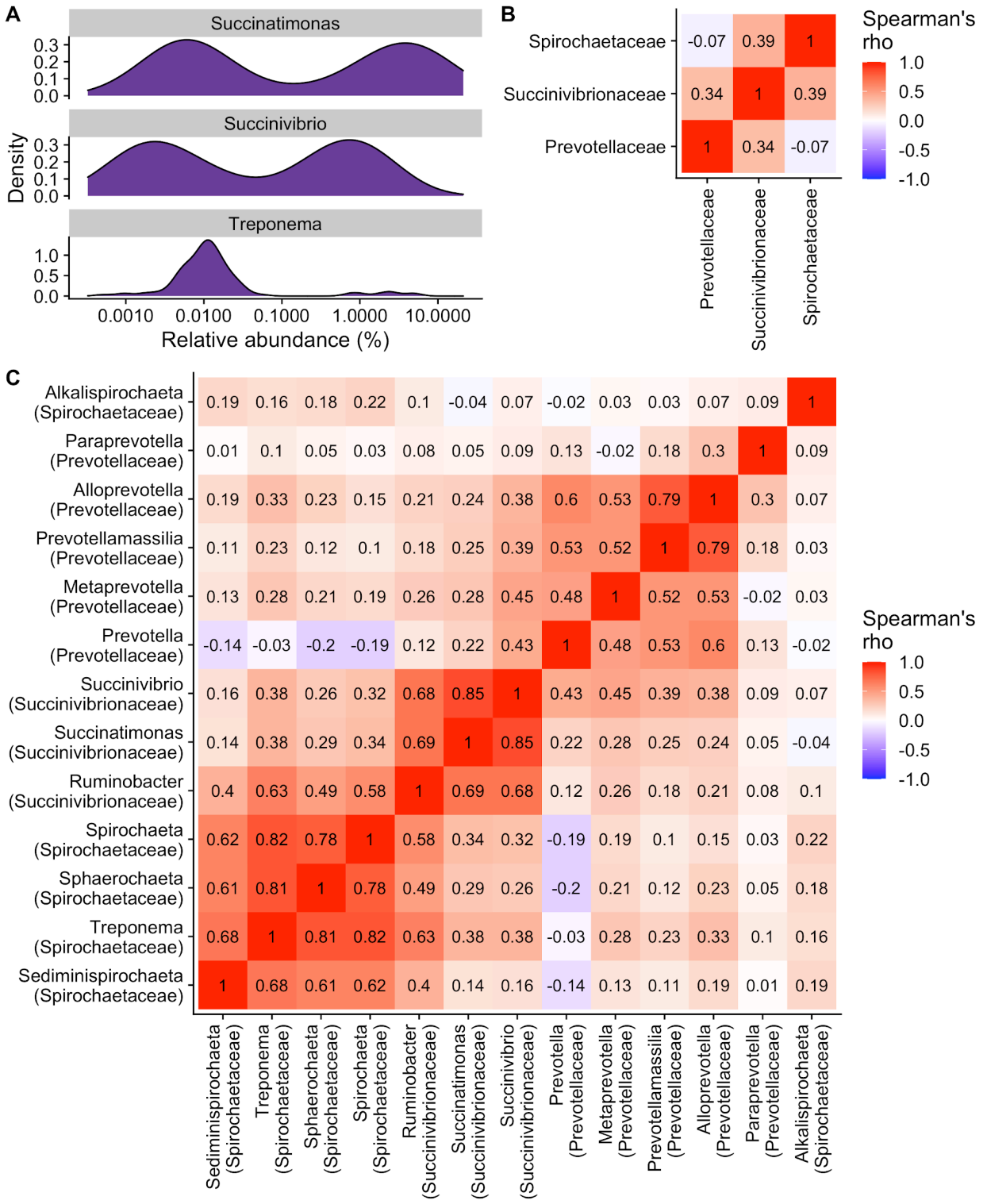
Supplementary Figure 1. Most abundant species and genera

Most abundant taxa by mean relative abundance shown for samples from two South African cohorts.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

649 (A) *Prevotella*, *Bacteroides,* and *Faecalibacterium* are the most abundant genera across both
650 study sites.
651 (B) The most abundant species are *Prevotella copri*, *Faecalibacterium prausnitzii,* and a
652 bacterium from the family Ruminococcaceae.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



653 **Supplementary Figure 2. Bimodal distribution of three VANISH taxa**

654 (A) *Succinatimonas, Succinivibrio*, and *Treponema* relative abundance values follow a bimodal

655 distribution in Bushbuckridge.

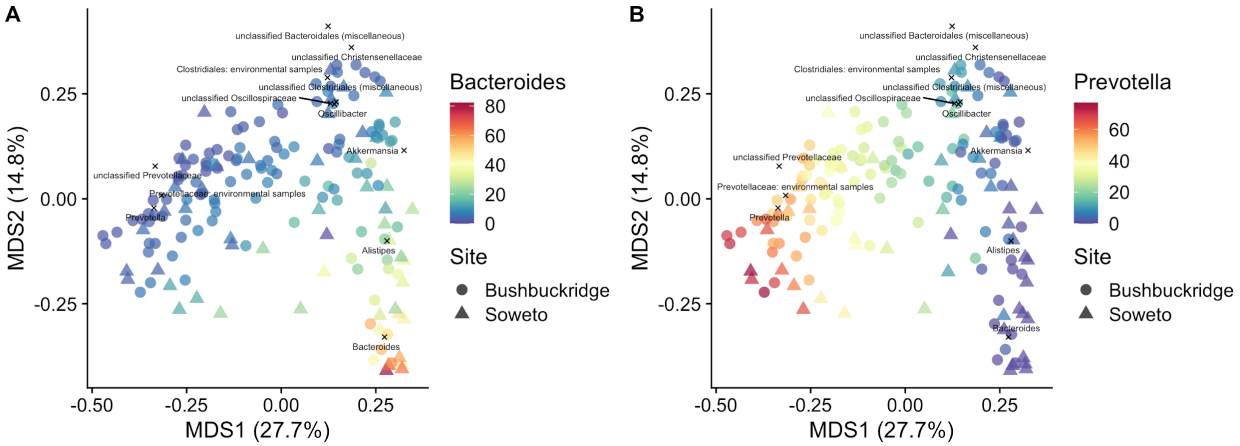
656 Across all South African samples, several VANISH families (B) and genera (C) are correlated,

657 with the exception of *Prevotella* and genera of the family *Spirochaetaceae* which are not

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

658 correlated with *Prevotella* (*Treponema*) or weakly anti-correlated with *Prevotella* (*Spirochaeta*,
659 *Sphaerochaeta*, *Sediminispirochaeta*).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



660 **Supplementary Figure 3. Bacteroides/Prevotella gradient across study population**

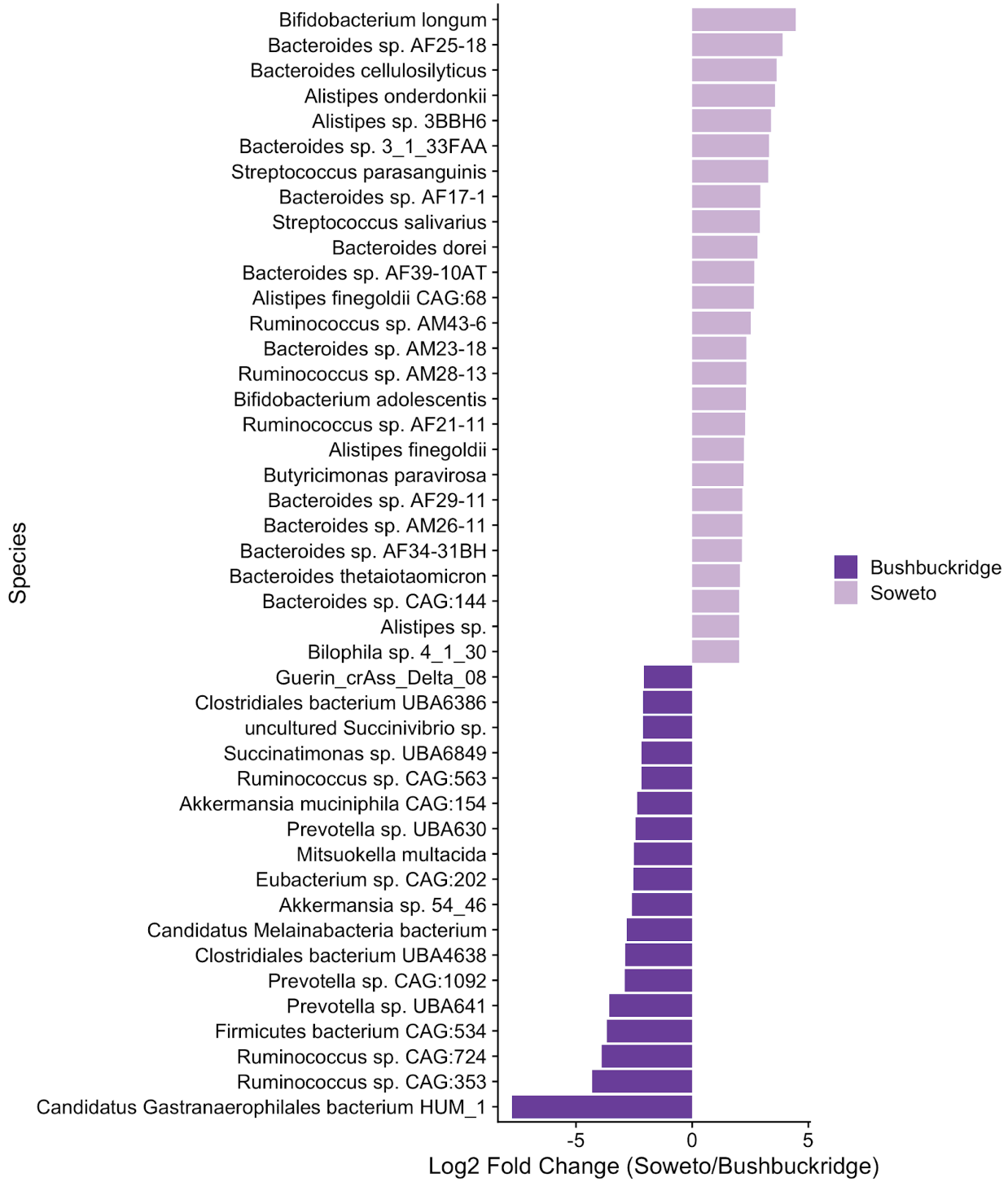
661 Multidimensional scaling ordination of species classifications in South African microbiome

662 samples colored by relative abundance of the genera *Bacteroides* (A) and *Prevotella* (B).

663 *Bacteroides* and *Prevotella* are major axes of variation across study samples. Text annotations on

664 MDS plot represent points with the highest relative abundance of the labeled genus.

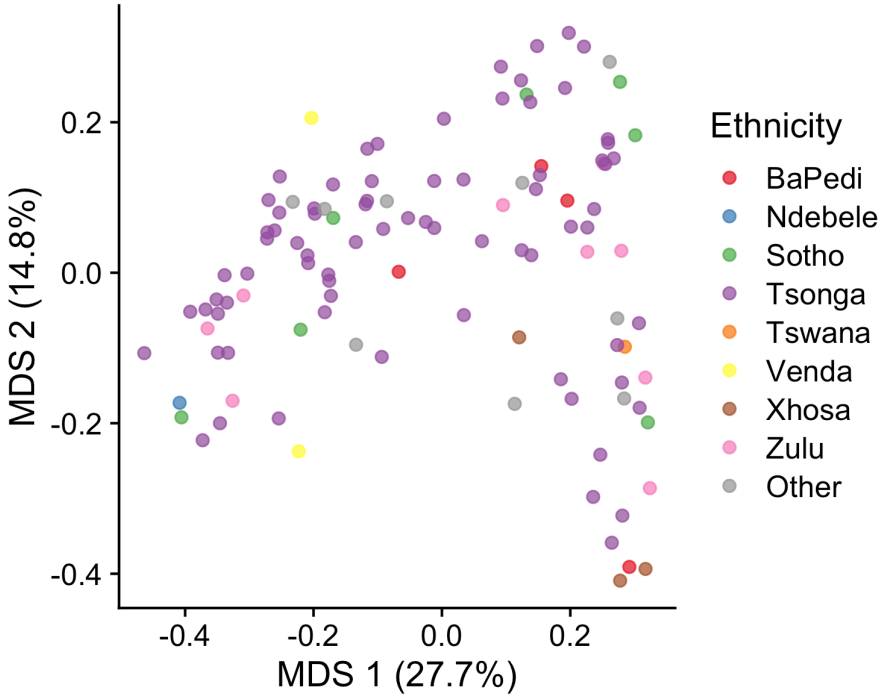
bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Supplementary Figure 4. Differentially abundant species between Bushbuckridge and Soweto

Differentially abundant microbial genera between rural Bushbuckridge and urban Soweto samples identified by DESeq2. Features with log2 fold change greater than one are shown (full results in Table S5).

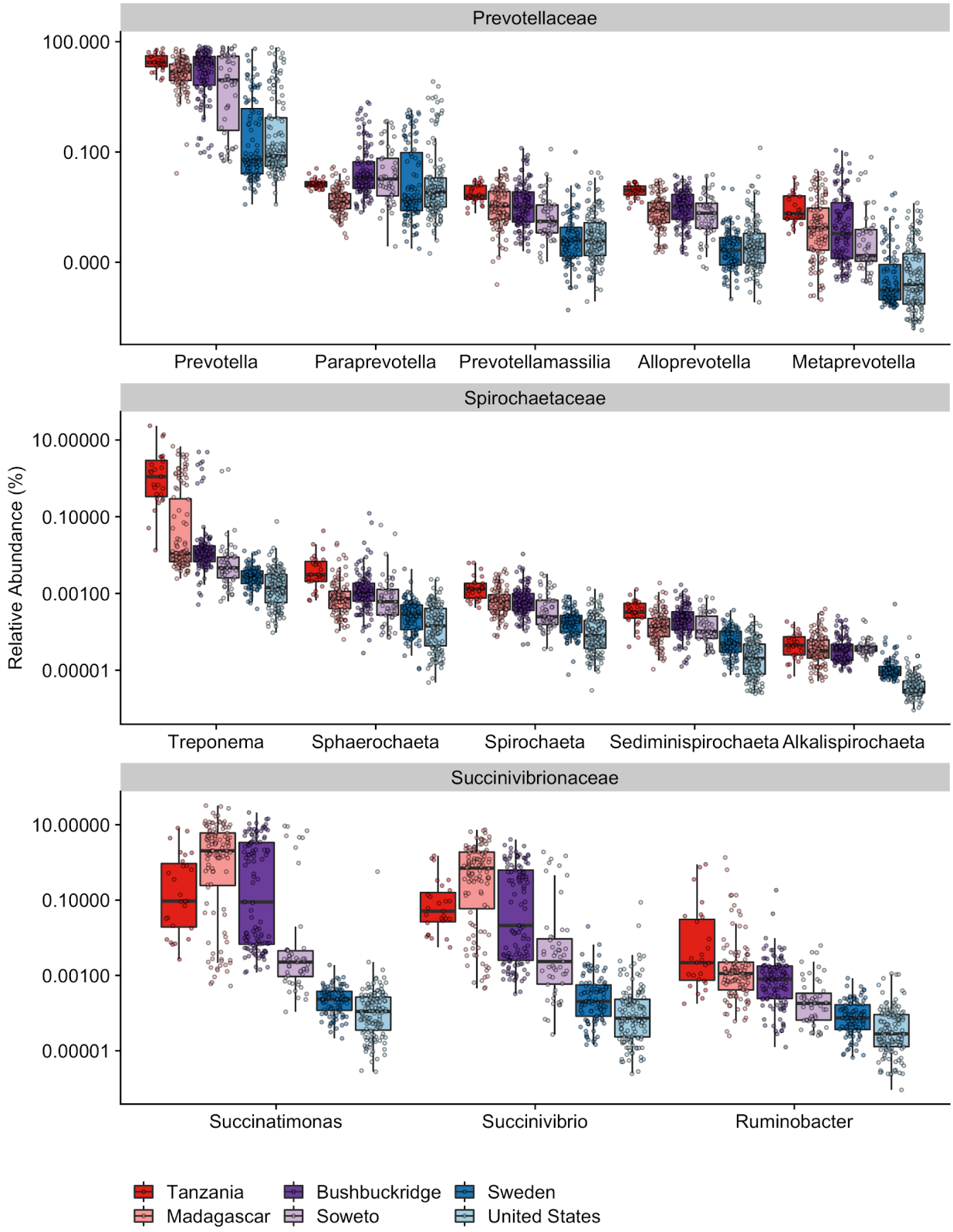
bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Supplementary Figure 5. South African microbiomes do not cluster by self-reported ethnicity

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



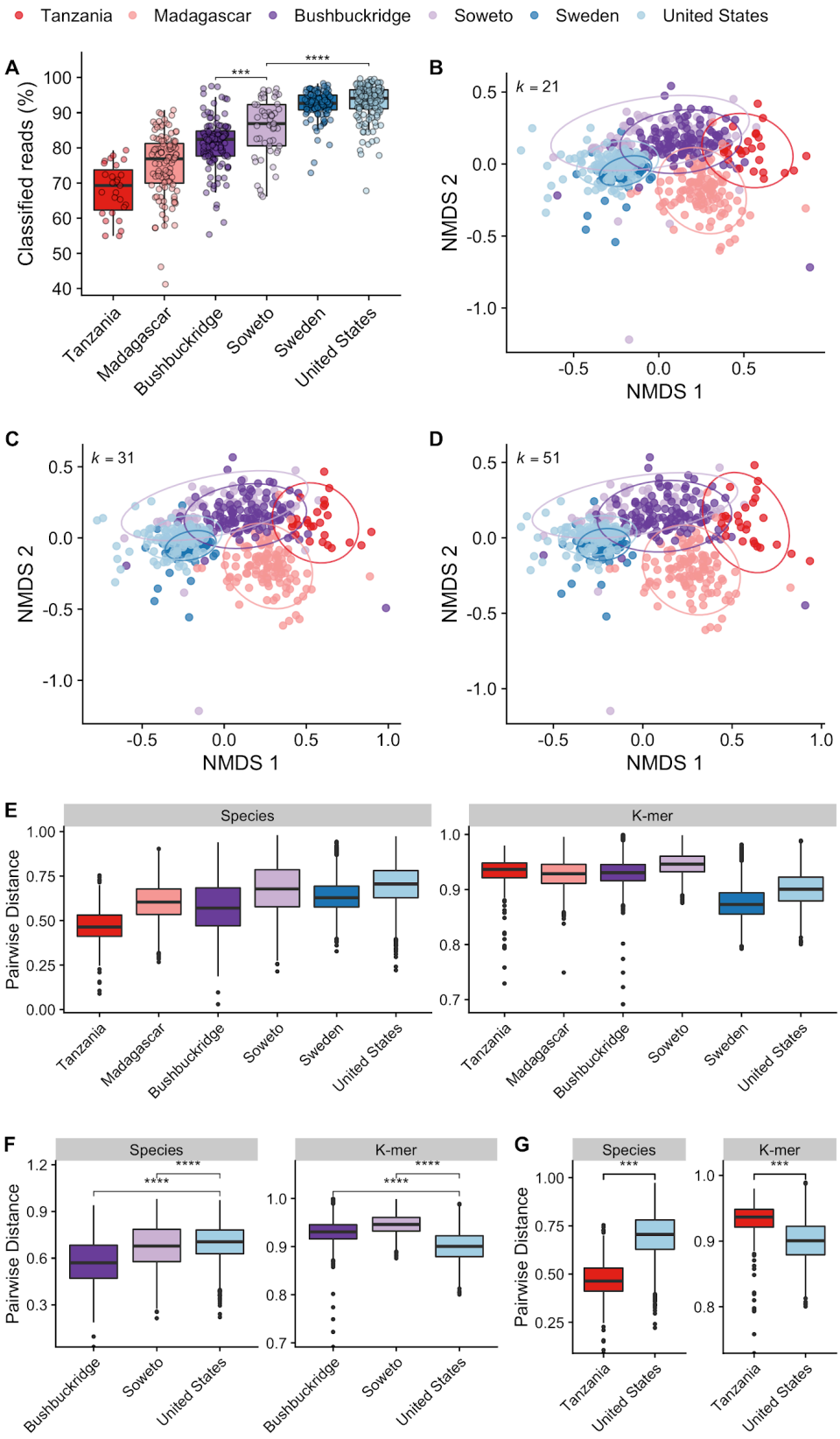
Supplementary Figure 6. Relative abundance of VANISH taxa in global cohort

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

677 abundance normalization in order to plot on a log scale. Relative abundance of most genera
678 trends toward decrease from nonwestern cohorts to western cohorts.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



679 **Supplementary Figure 7. Comparison of beta diversity between communities calculated by**

680 **taxonomy versus nucleotide *k*-mer composition**

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

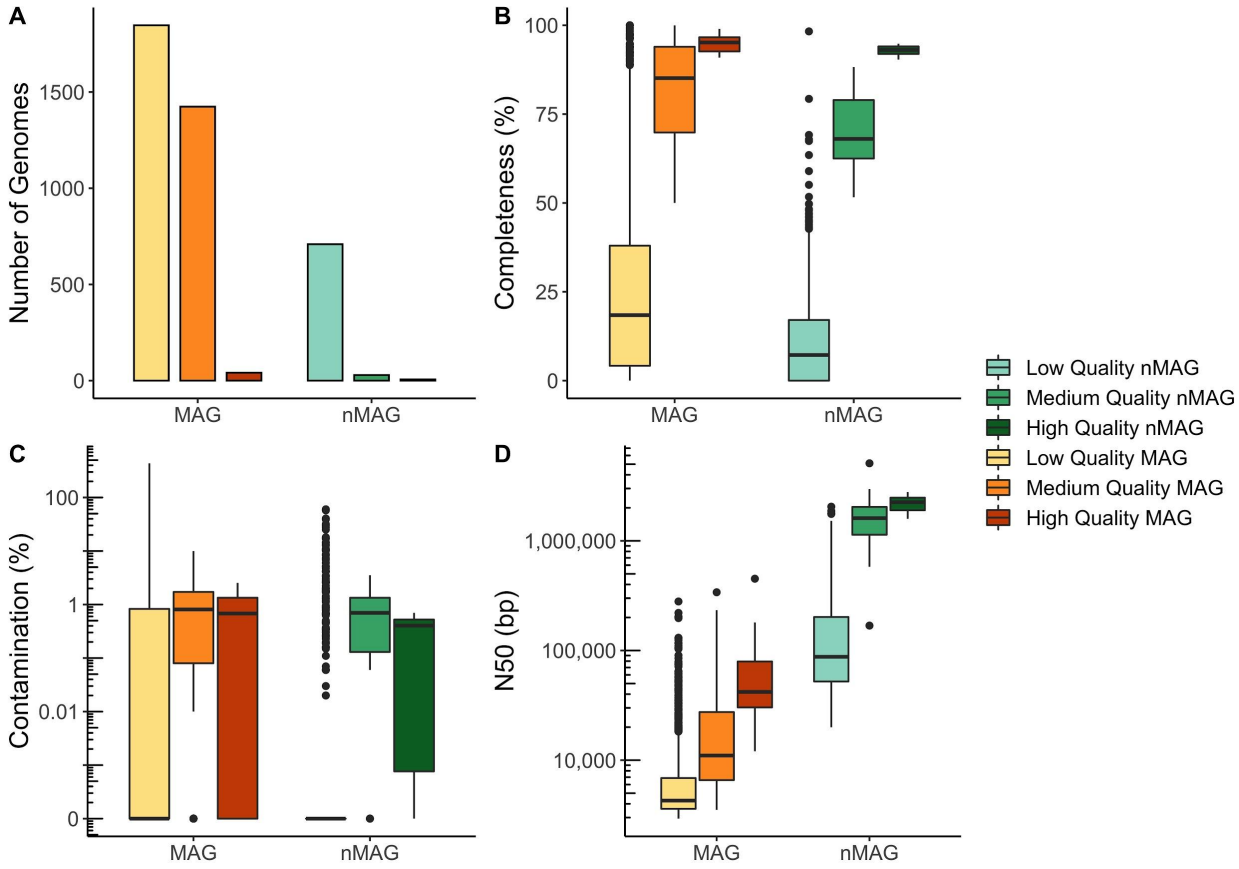
(A) Percentage of reads classified at any taxonomic rank per cohort based on a reference database of all scaffold or higher quality reference genomes in GenBank and RefSeq as of January 2020. Western microbiomes have a higher percentage of classifiable reads compared to nonwestern microbiomes (Wilcoxon rank-sum test $p < 0.001$).

(B-C) Nucleotide sequences of microbiome sequencing reads were compared using MinHash sketches. This reference-free approach is not constrained by comparison to existing genomes and therefore allows direct comparison of sequences. Briefly, a hash function generates signatures at varying sequence lengths (k) and hashes can be compared between samples. Distances shown are non-metric multidimensional scaling (NMDS) of Jaccard distance between samples based on shared k-mer composition. Data shown here are generated from comparisons at k=21 (approximate genus-level, B), k=31 (approx. species-level, C), and k=51 (approx. strain-level, D) k=31 (Koslicki and Falush, 2016). Global microbiomes cluster distinctly by cohort.

(E-G) Comparison of pairwise beta diversity within communities assessed by Bray-Curtis distance based on species-level classifications and Jaccard similarity of nucleotide *k*-mer sketches. (E) All populations. (F) South African populations (Bushbuckridge and Soweto) compared to the United States cohort. Beta diversity measured by Bray-Curtis distance is higher in Soweto but lower in Bushbuckridge compared to the United States. However, reference-independent *k*-mer comparisons indicate that nucleotide dissimilarity is higher within both South African populations compared to the United States. (G) Species-based Bray-Curtis distance indicates that there is more beta diversity within the United States cohort, but *k*-mer distance indicates an opposite pattern.

(A, E-G) Wilcoxon rank-sum test, significance values denoted as follows: (***) $p < 0.001$, (****) $p < 0.0001$.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Supplementary Figure 8. Summary statistics for Illumina and nanopore MAGs generated from all samples

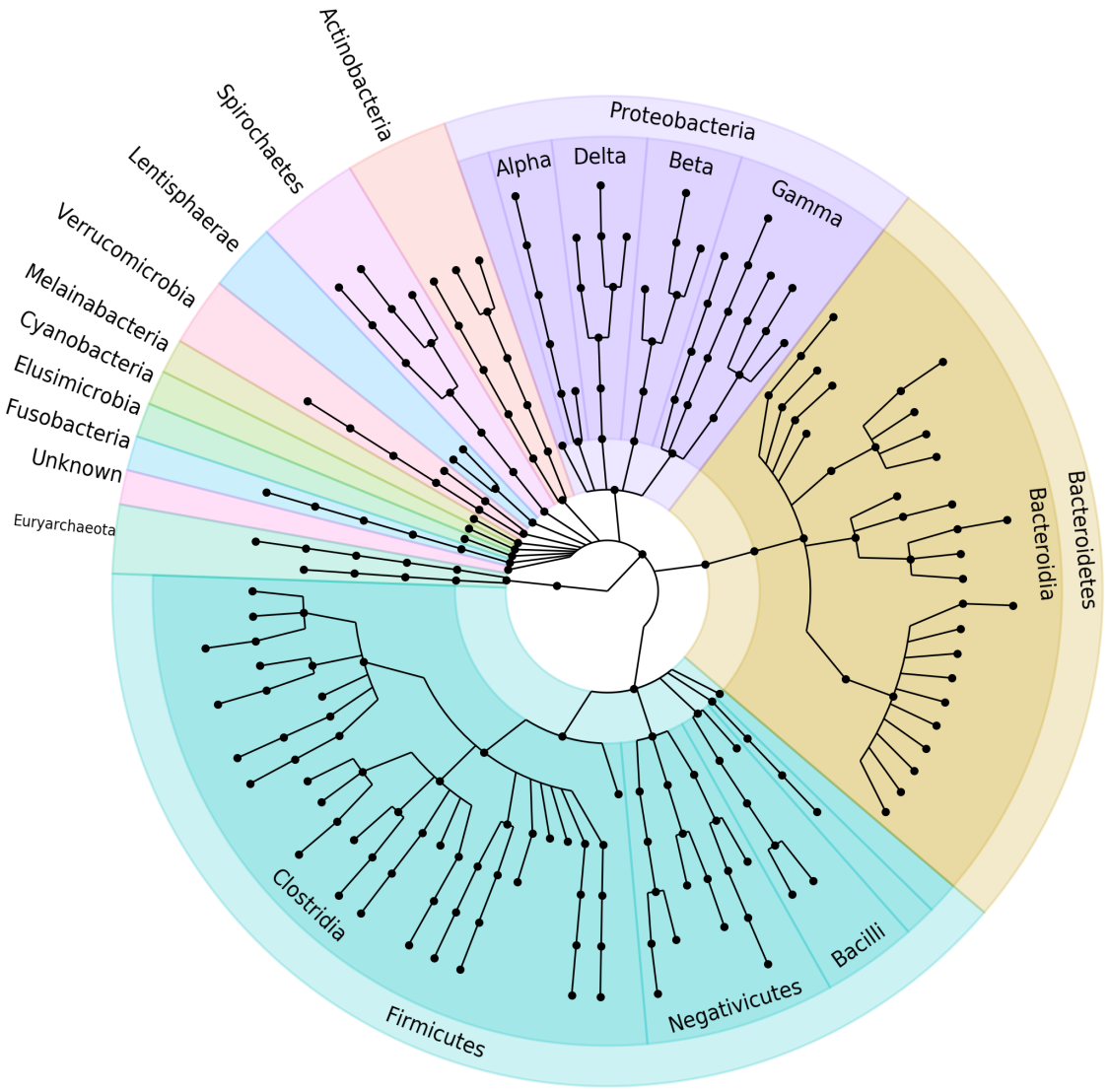
(A) Number of low-, medium-, and high-quality genomes as evaluated with Bowers et al. standards

(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.

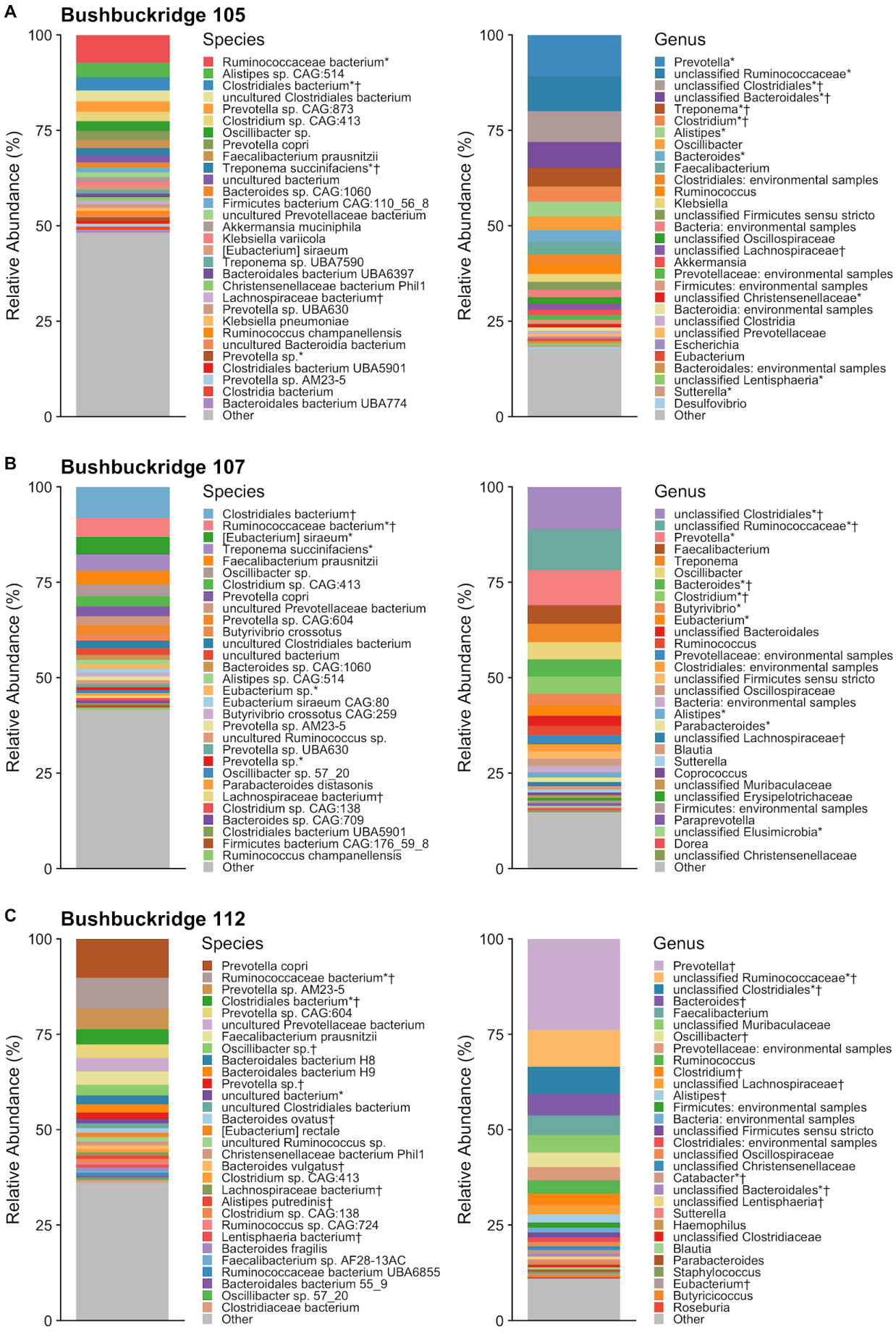
bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Supplementary Figure 9. Taxonomy of dereplicated Illumina MAGs from all samples

Taxonomic classification of dereplicated medium- and high-quality Illumina MAGs. Multiple MAGs at the same classification level are collapsed into single points.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

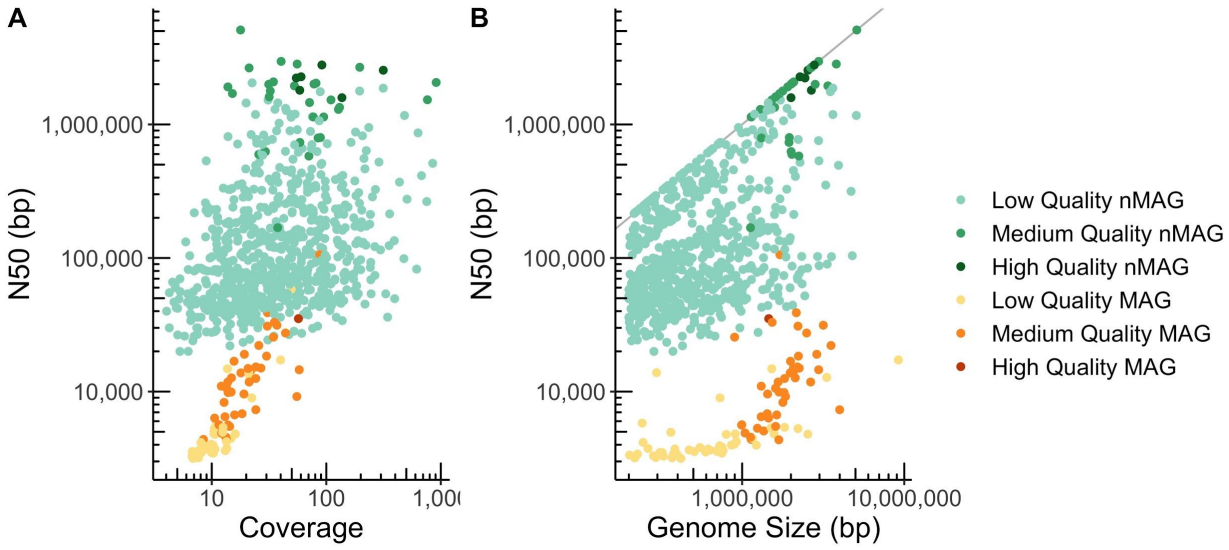


bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Figure 10. Taxonomic composition for samples selected for nanopore sequencing

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

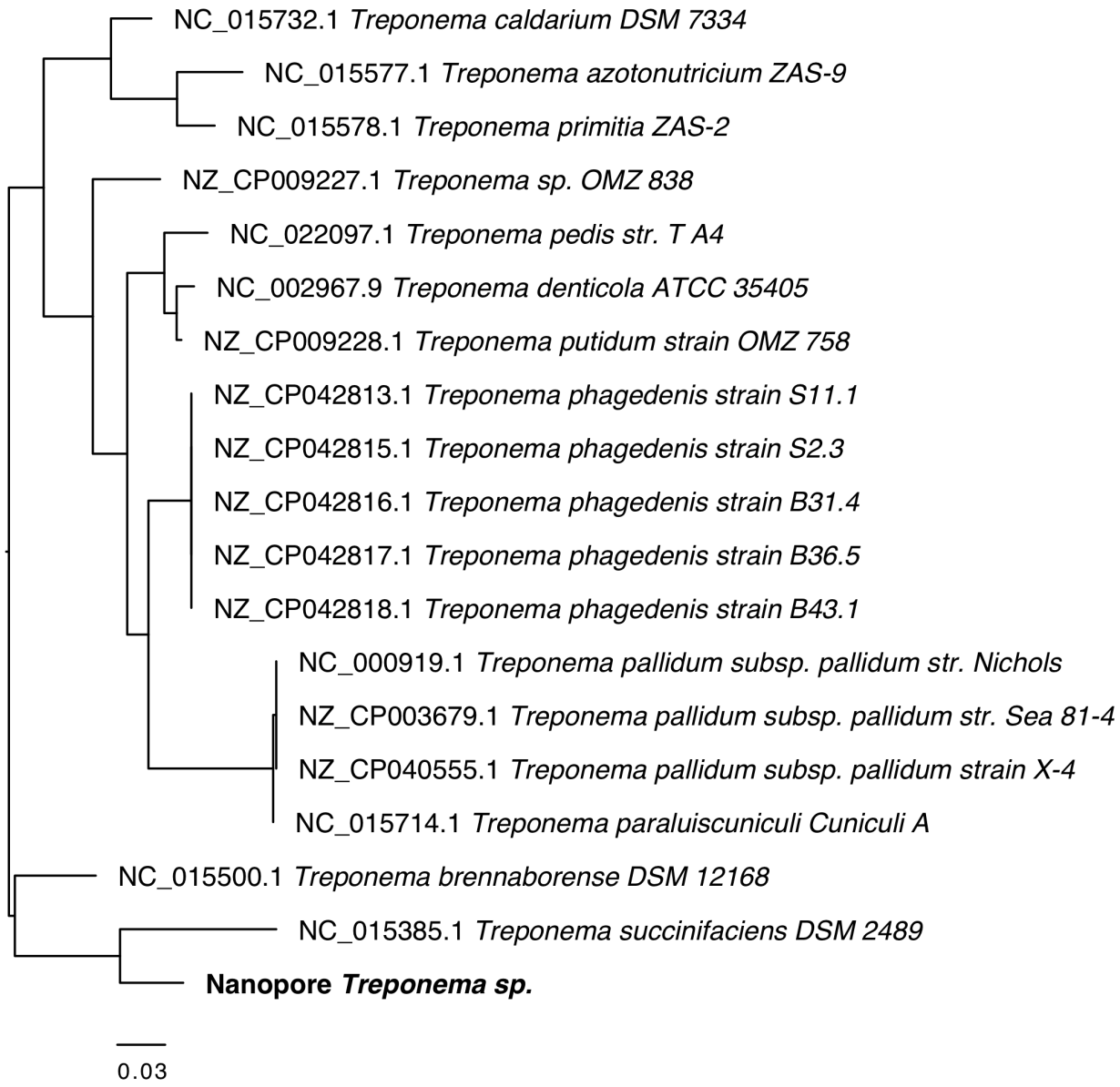
bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Supplementary Figure 11. 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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



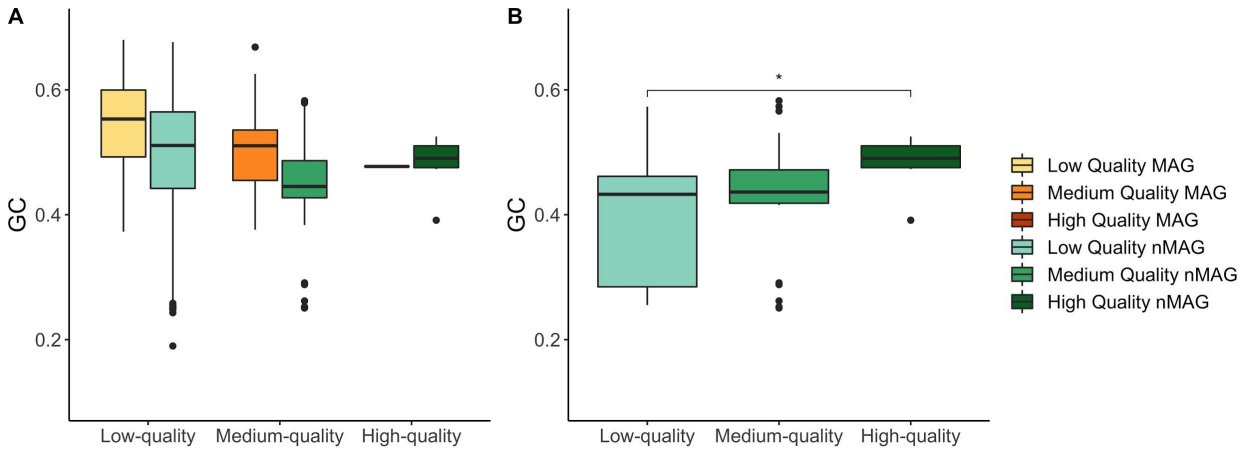
726 **Supplementary Figure 12. Phylogeny of *Treponema* 16S rRNA sequences**

727 Phylogeny of 16S rRNA sequences from species of the genus *Treponema* show that the

728 *Treponema* sp. assembled via Nanopore sequencing is most related to *T. succinifaciens*, but is

729 phylogenetically distinct.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

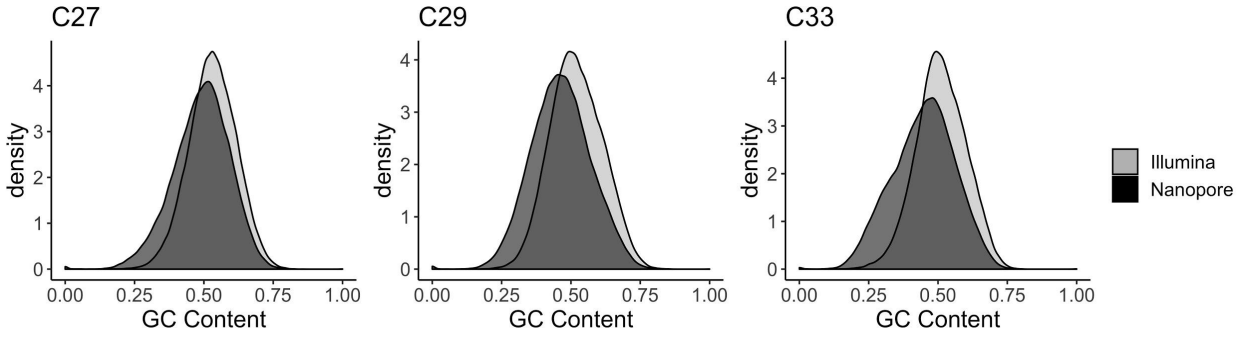


730 **Supplementary Figure 13. GC content of MAGs and nMAGs generated from three**
731 **Bushbuckridge samples**

732 (A) GC content range of MAGs and nMAGs.

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Supplementary Figure 14. GC content of nanopore and Illumina sequencing reads generated from three Bushbuckridge samples

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

741 **Supplementary Figure 15. Phylogeny of Mollicutes 16S rRNA sequences**

742 Phylogeny of 16S rRNA sequences from species of the class Mollicutes showing the Mollicutes

743 and Mycoplasma genomes assembled via nanopore sequencing.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

744 Ahlmann-Eltze, C. (2019). ggsignif: Significance Brackets for "ggplot2."

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

749 Almeida, A., Mitchell, A.L., Boland, M., Forster, S.C., Gloor, G.B., Tarkowska, A., Lawley,
750 T.D., and Finn, R.D. (2019). A new genomic blueprint of the human gut microbiota. Nature *568*,
751 499–504.

752 Angelakis, E., Bachar, D., Yasir, M., Musso, D., Djossou, F., Gaborit, B., Brah, S., Diallo, A.,
753 Ndombe, G.M., Mediannikov, O., et al. (2019). Treponema species enrich the gut microbiota of
754 traditional rural populations but are absent from urban individuals. New Microbes New Infect
755 *27*, 14–21.

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

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

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

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

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

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

773 Brown, C.T., and Irber, L. (2016). sourmash: a library for MinHash sketching of DNA. JOSS *1*,
774 27.

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

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

786 De Filippo, C., Cavalieri, D., Di Paola, M., Ramazzotti, M., Poullet, J.B., Massart, S., Collini, S.,
787 Pieraccini, G., and Lionetti, P. (2010). Impact of diet in shaping gut microbiota revealed by a
788 comparative study in children from Europe and rural Africa. Proc. Natl. Acad. Sci. U. S. A. *107*,
789 14691–14696.

790 Di Rienzi, S.C., Sharon, I., Wrighton, K.C., Koren, O., Hug, L.A., Thomas, B.C., Goodrich, J.K.,
791 Bell, J.T., Spector, T.D., Banfield, J.F., et al. (2013). The human gut and groundwater harbor
792 non-photosynthetic bacteria belonging to a new candidate phylum sibling to Cyanobacteria. Elife
793 *2*, e01102.

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

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

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

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

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

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

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

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

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

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

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

828 Hunt, M., Silva, N.D., Otto, T.D., Parkhill, J., Keane, J.A., and Harris, S.R. (2015). Circlator:
829 automated circularization of genome assemblies using long sequencing reads. Genome Biol. *16*,
830 294.

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

834 Kabudula, C.W., Houle, B., Collinson, M.A., Kahn, K., Gómez-Olivé, F.X., Clark, S.J., and
835 Tollman, S. (2017a). Progression of the epidemiological transition in a rural South African
836 setting: findings from population surveillance in Agincourt, 1993--2013. BMC Public Health *17*,
837 424.

838 Kabudula, C.W., Houle, B., Collinson, M.A., Kahn, K., Gómez-Olivé, F.X., Tollman, S., and
839 Clark, S.J. (2017b). Socioeconomic differences in mortality in the antiretroviral therapy era in
840 Agincourt, rural South Africa, 2001-13: a population surveillance analysis. Lancet Glob Health
841 *5*, e924–e935.

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

844 Kassambara, A. (2020). ggpubr: "ggplot2" Based Publication Ready Plots.

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

847 Krueger, F. Trim Galore!

848 Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg,
849 S.L. (2004). Versatile and open software for comparing large genomes. Genome Biol. *5*, R12.

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

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

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

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

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

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

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

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

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

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

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

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

926 Santosa, A., and Byass, P. (2016). Diverse Empirical Evidence on Epidemiological Transition in
927 Low- and Middle-Income Countries: Population-Based Findings from INDEPTH Network Data.
928 PLoS One *11*, e0155753.

929 Sato, M.P., Ogura, Y., Nakamura, K., Nishida, R., Gotoh, Y., Hayashi, M., Hisatsune, J., Sugai,

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

M., Takehiko, I., and Hayashi, T. (2019). Comparison of the sequencing bias of currently available library preparation kits for Illumina sequencing of bacterial genomes and metagenomes. DNA Res. *26*, 391–398.

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

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

Scholz, M., Ward, D.V., Pasolli, E., Tolio, T., Zolfo, M., Asnicar, F., Truong, D.T., Tett, A., Morrow, A.L., and Segata, N. (2016). Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nat. Methods *13*, 435–438.

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

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

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

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

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

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

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

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

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

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

Tyler, A.D., Mataseje, L., Urfano, C.J., Schmidt, L., Antonation, K.S., Mulvey, M.R., and

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.18.099820; this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

967 Corbett, C.R. (2018). Evaluation of Oxford Nanopore's MinION Sequencing Device for
968 Microbial Whole Genome Sequencing Applications. Sci. Rep. *8*, 10931.

969 Vangay, P., Johnson, A.J., Ward, T.L., Al-Ghalith, G.A., Shields-Cutler, R.R., Hillmann, B.M.,
970 Lucas, S.K., Beura, L.K., Thompson, E.A., Till, L.M., et al. (2018). US Immigration Westernizes
971 the Human Gut Microbiome. Cell *175*, 962–972.e10.

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

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

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

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

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

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

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

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

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

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