Population Stratification in the Gut Microbiota of Bali is Associated with Transitional Lifestyle

Clarissa Asha Febinia
Eijkman Institute for Molecular Biology https://orcid.org/0000-0002-6918-8529
Safarina G. Malik
Eijkman Institute for Molecular Biology
Ratna Djuwita
Universitas Indonesia
I Wayan Weta
Universitas Udayana Fakultas Kedokteran
Desak Made Wihandani
Universitas Udayana Fakultas Kedokteran
Rizka Maulida
Universitas Indonesia
Herawati Sudoyo
Eijkman Institute for Molecular Biology
Andrew J. Holmes (andrew.holmes@sydney.edu.au)
University of Sydney https://orcid.org/0000-0002-7406-4387

Research

Keywords: gut microbiota, enterotypes, Bali, Indonesia, lifestyle transition, nutrition, population stratification

Posted Date: July 16th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-40341/v1

Abstract

Background: Human living conditions, such as food availability and the built environment, contribute to environmental forces that influence gut microbiota composition. Understanding the impact of the environment on microbiota assembly and its association with human health has multiple potential applications. Indonesia is a densely populated country that has been undergoing a dramatic societal change for the past two decades. It is distinctive in that it occupies an archipelago that imposes diverse geographic and cultural boundaries. The relationship between diet, microbiota, and health is poorly known in Indonesians and represents a natural study for the interaction between ethnogeographic factors and nutrition in microbiota assembly.

Results: Here we show the first comprehensive report of the gut microbiota in adults from Bali, Indonesia (n=41). Their microbiotas clustered into two distinct community types: a *Prevotella*-rich (Type-P) and a *Bacteroides*-rich (Type-B) community. The Type-P individuals had lower alpha diversity (p <0.001, Shannon) and more incidence of obesity. The two community types are significantly different in their inter-genus co-abundance pattern (p <0.001, ANOSIM, Wilcoxon test). Further analyses with diet and obesity data showed that the presence of two distinct community types in Bali is a significant confounder for identifying health markers. In a multi-country dataset (n=257), the Bali microbiota indicates a transitional state from a subsistent (*Prevotella*-dominant) to industrial (*Bacteroides*-dominant) society. The two largest axes in a Principal Coordinate Analysis of weighted UniFrac distance explained the majority of variance between samples across countries (49.1%). Microbial dissimilarity across populations is significantly associated with *Prevotella* and *Bacteriodes* abundance (p <0.001, Generalized Additive Model).

Conclusion: Our data showed that lifestyle transitions have a strong influence on the frequency of microbiota community types in a population. The Bali microbiota is undergoing a shift towards a *Bacteroides*-dominant community which reflects the ongoing transition of nutrition, socio-economy, and lifestyle the society. Although enterotypes obscured the detection of health markers, our findings collectively suggest that enterotypes may be useful in future studies for informing population-level stratification in large heterogenetic datasets.

Introduction

The epidemic of chronic immuno-metabolic diseases, such as obesity and diabetes, presents a major challenge for health systems around the world. Epidemiological studies show that the increased incidence has occurred in association with environmental changes due to modernization, including more sedentary lifestyle and nutritional changes. They further show that for many of these nutrition-related chronic diseases (NRCD), differences in the human gut microbiota are also associated with them. For many NRCD, mechanistic links between microbes and pathological processes have been identified in animal models. This complex aetiology is known as *dysbiosis* – an undesirable state of the body system in which microbiota factors participate in the pathophysiological processes of the host. Two inter-related

public health questions thus arise. Do environmental changes drive the microbiota in ways that globally alter the risk of dysbioses within a population? If so, which dimensions of the environment are most significant?

Cross-sectional studies of human populations occupying distinct environments have repeatedly found significant differences in the human gut microbiota [1–9]. It is widely accepted that these differences are potentially relevant to human health, however, neither the environmental factors driving microbiota differences nor their relevance for public health is well understood. A reason for this uncertainty is that human populations vary in multiple ways, including genetic, cultural, and geographic factors, each of which can also interact over time. A good example is the human diet. An individual's food intake is shaped by both cultural and geographic factors that collectively define their nutrient environment.

The term nutrition transition has been used to describe the net process of a society moving from a foraging or subsistent agriculture lifestyle to a modern urbanized lifestyle with industrialized food supply chain [10]. Differences in the structure of the gut microbiotas are typically seen in people from societies at different points of the nutrition transition [1–9]. However, separating the effects of dietary change from other factors (e.g. socioeconomic status, healthcare, and ethnicity) is difficult. Available data can be roughly categorized into three groups: 1) studies which compared pre- and post-transition societies at a specific point in time [1–7, 9, 11, 12]; 2) longitudinal study of demographic groups as they immigrate from a pre-transition to post-transition society [8]; and 3) experimental diet studies in individuals [13, 14]. A recurrent finding in these cross-sectional studies is that some taxa, notably *Prevotella* and *Bacteroides*, are differentially represented in microbiotas of pre- and post-transition populations.

The clustering of human gut microbial communities by *Prevotella* and *Bacteroides* was first reported by Arumugam *et al*. They coined the term enterotypes – which also included a third grouping based on *Ruminococcus* – and cautiously proposed that they would have applications in diagnostics [15]. A caveat in defining enterotypes, which was raised in that study and has been elaborated since by multiple authors, is that it is a complex process in which their observation is a product of analytical methods that are dataset dependent [16, 17]. Nevertheless, the *Prevotella*-type community is repeatedly found to be more prevalent in subsistent societies, and the *Bacteroides*-type in industrial societies. Such findings have been used to suggest that diet drives these microbiota variations [5, 14, 17]; and differences in intake of dietary fibre and animal source foods, in particular, have been postulated as key factors [5, 11, 13, 14, 18].

While numerous studies have shown that diet can induce a rapid change in the microbiota of individuals [13, 14, 19], evidence that it drives enterotype formation is lacking. Long-term diet patterns have been associated with enterotypes [14, 20–22], but other societal factors influencing microbial dispersal and community assembly are also associated. These barriers include, but are not limited to, water supply, housing density, birth mode, child-rearing, and infection control in healthcare settings. Effects on dispersal would almost certainly intersect with diet changes to influence microbiota assembly.

In this study, we aimed to explore the gut microbiota of a demographically narrow group of individuals, raised during a period of rapid economic and nutrition transition [23, 24]. In this cohort of 41 young

females from Bali, Indonesia, we found two community types in an approximately equal frequency. Further, we explored the association of these community types with other aspects of microbial structural composition, diet, and obesity. Lastly, we look for patterns of community type distribution across different human populations by conducting a meta-analysis that incorporated publicly available data from 216 individuals from five populations with distinct socio-cultural and geographical backgrounds. We showed that the Bali cohort has a distinct pattern of community type frequency and that it is not explained by contemporary diet.

Methods

Study design

This is a cross-sectional study of the gut microbiota of Bali individuals from Indonesia in comparison to individuals from other populations. The aim is to identify microbiota patterns associated with lifestyle transition. Particularly, we looked for microbiota features in pre- and post-transition societies; and further, explored the association of the differences with nutritional intake and obesity.

Enrolment of Bali individuals

A cross-sectional sampling was performed. Study enrolment for the Bali cohort was conducted on 19–23 January 2015, at the Faculty of Medicine, Udayana University, Denpasar, Bali Province, Indonesia. We recruited all available volunteers in the area at the time of enrolment. We received written and signed informed consent from all participants. A clinician declared the volunteers were not suffering from chronic or infectious diseases at the time of enrolment. The volunteers self-declared that they had not consumed antibiotics, underwent surgery, or became pregnant within 3 months before sample collection.

Anthropometry and diet data collection

Age was determined by date of birth. Weight and height were measured without shoes to the nearest 0.5 cm and 0.1 kg, respectively. BMI was calculated as body weight by squared height (kg/m2). Dietary information was collected using 24-hour food recall and food frequency questionnaires. NutriSurvey software (2007 English version of EBISpro, Germany) was used for analysis of nutrient content. The analysis used the Indonesian food database developed by The Southeast Asian Ministers of Education Organization – Regional Tropical Medicine and Public Health Network (SEAMEO-TROPMED) and the Deutsche Gesellschaft für Technische Zusammenarbeit (GTZ) GmbH Federal Republic of Germany [25].

Stool samples

Faecal samples were self-collected by the volunteers in a pre-sterilized pot. Within 4 hours of defecation, the samples were put in -20 ^{o}C freezer for 8–48 hours. Frozen samples were packed into aero-thermal insulated containers with freezer packs and delivered to the Eijkman Institute for Molecular Biology in Jakarta within 5 hours and stored in -20 ^{o}C freezer on arrival. Within 48 hours of delivery, we extracted metagenomic DNA using the PowerSoil DNA Extraction Kit (MOBIO Laboratories Inc., USA, Catalogue No.

12888-100) following the manufacturer's protocol, but sample homogenization was done for 15 minutes using Vortex Genie-II (MOBIO Laboratories Inc., USA). DNA quality was assessed using NanoDrop® Spectrophotometer (Thermo Fisher Scientific) and 1% agarose gel electrophoresis. Extractions for low quality DNA samples were repeated. Good quality DNA samples were stored in -20 ^{o}C before being sent for sequencing at the Ramaciotti Centre for Genomics at the University of New South Wales, Australia, in May 2015.

Sequence processing

We sequenced the variable region 4 (V4) of 16S rRNA gene from the extracted metagenome, using 515F/806R sequencing primers, as previously described [26]. Paired-end sequence reads were merged *in silico* using the fastq-join tool (parameters: minimum overlap of 8 bp and maximum difference percentage of 10%) [27]. Merged sequences that contained low-quality reads (Q < 30) were truncated. We selected reads with lengths between 230 bp and 255 bp. Sequence data was imported into Quantitative Insights Into Microbial Ecology (QIIME) version 1.9.1 [28]. Chimera sequences were removed using *uchime* [29]. Operational Taxonomic Units (OTUs) were picked at 97% sequence similarity according to *uclust* method. Reads that did not align to the greengenes reference database (version 13.8) were retained as *de novo* OTU*s* [30]. Singleton OTUs were removed. We assigned taxonomy using Ribosomal Database Project Naïve Bayesian Classifier (RDP Classifier) at 80% similarity to Greengenes (version 13.8) reference dataset [31].

Diversity analysis

Microbial alpha-diversity was computed using Faith's Phylogenetic Diversity [32], Chao1, and Shannon index. Distances between samples based on the microbiota (beta diversity) were calculated as weighted and unweighted UniFrac distance [33] and analysed using Principal Coordinate Analysis (PCoA) [34]. We performed a clustering analysis on the Bali dataset according to Ward method on the weighted UniFrac distance [35].

Co-Abundance Groups (CAG) analysis

To identify fine-scale taxonomy patterns in the Bali dataset, we performed co-abundance network analysis using genus-level data at the sequencing depth of 90,000 sequences per sample. Only genera with a presence in more than 30% of sample size (12 samples) were used. We assessed pairwise relationships of these genera using Kendall correlation test (t). Positive Kendall scores ($0 < t < 1$) were taken as co-existing relationships, while negative scores ($-1 < t < 0$) were taken as contra-existing relationships. The scores were gathered into a matrix of t-scores, mathematically inversed ($1 - t$), and put through a hierarchical clustering analysis according to Ward's method [10]. We performed empirical clustering configurations using $k = 2$ to $k = 12$ and tested the cluster distinction in these configurations using Analysis of Similarity (ANOSIM) with 10,000 permutations. The highest configuration which passed the test ($p < 0.05$) was determined as co-abundance groups (CAG).

Comparative dataset

We downloaded merged and filtered V4 pair-ended sequence of 16S rRNA gene from the Metagenomics RAST server (MG-RAST) public repository [36]. Sample selection used the following inclusion criteria: study design (observational), sample type (stool), analysis region of the 16S rRNA gene (V4), and age group (adults aged 18–60). We obtained samples from Hadza hunter-gatherers, Malawians, Guahibo Amerindians, Italians, and Americans (Project No. mgp401 and mgp7058). Metadata for the reference dataset, including data on geography, age, and sex, were obtained from their original studies [2, 3]. OTUs were assigned *de novo* at 97% sequence similarity. OTUs with < 0.005% presence within a population were removed. After the taxonomy assignment, the OTUs were grouped at species-level (Greengenes Level 7). The resulting table was rarefied to 4,000 sequences per sample and used for calculating inter-sample weighted UniFrac distance. The Greengenes reference sequence (version 13.8) served as the base for calculating phylogenetic relationships between the species-level OTUs.

Data analyses

All tests and figures were generated in R Project for Statistical Computing version 3.4.1 (www.rstudio.com). Images were generated in RStudio and compiled in Inkscape version 0.48. The normality of data distributions was tested with Shapiro-Wilk test. Differences in continuous variables between sample groups were tested with Student's T test if the distribution was normal, or Wilcoxon's Rank-Sum test if the distribution was skewed. When multiple comparisons were performed, *p*-values were corrected using the False Discovery Rate (FDR) method [37]. Significant correlations are determined as having a $p < 0.05$, or $q < 0.05$ if FDR adjustments were performed. Beta diversity data was visualised using Principal Coordinate Analysis (PCoA). Comparison between groups in weighted UniFrac data used ANOSIM. Associations between bacterial genus and beta diversity were analysed using vector regression fitting and Permutational Multivariate Analysis of Variance (PERMANOVA) [38]. The projection of *Prevotella* and *Bacteroides* onto PCoA used Generalized Additive Model (GAM). Pairwise correlation relationships were measured using the Kendall Rank-Correlation test. Associations between CAGs, BMI, and macronutrient intake were analysed using univariate and multivariate rank-based regression model; and to assess the robustness of multivariate models, we used the reduction in dispersion test as previously described [39]. Models that passed the test ($p < 0.05$) are considered significant.

Results

Demographics and diet of the Bali cohort

We recruited 47 female individuals between 18 to 27 years old in Denpasar City, Indonesia. Faecal samples were obtained from 41 individuals, of whom 36 identified themselves as ethnically Balinese, while the other four individuals identified as Batak Toba, Javanese, Chinese descend, and a Javanese-Balinese admixture. Four individuals were not born in the area but have lived in Bali for at least 2 years and they were born in cities (Jakarta, Bekasi, and Pangururan). There were six (14%) obese individuals with body mass index (BMI) over 30 kg/m^2 (Table 1).

Food recall and food frequency questionnaires revealed limited dietary variation. Energy intake was low (Mean ± SD = 1,359 ± 599 kcal daily), but all lay within the Acceptable Macronutrient Distribution Ranges (AMDR). All individuals had rice as the staple source of energy. Rice is also the primary source of plant-based protein; followed by tofu and tempeh (Additional File 1: Fig. S1). Meat consumption was generally once or twice daily. Most individuals had either chicken, pork, or fish. Beef intake varied significantly, with 6 out of 10 people abstained from it. Three individuals were vegan.

Table 1
Anthropometric description of subjects

	All	**Lean**	**Obese**	***p***
Variable	*(n = 41)*	*(n = 35)*	*(n = 6)*	(lean vs. obese)
Age (years)	21 (20–21)	21 (20–21)	22 (20.5–22.8)	0.0697 [a]
Weight (kg)	53 (47.7–59.5)	50.3 (47.6–55.1)	90 (83.3–107.9)	**0.0001** [a]
Height (cm)	157.1 ± 4.9	156.6 ± 4.8	159.9 + 5.1	0.1851 [b]
BMI (kg/m^2)	21.1 (19.7–25.1)	20.4 (19.4–22.4)	35.8 (33.2–41.5)	**0.0001** [a]

Note: Values are shown as mean ± standard deviation and median (quantile 1 – quantile 3) for variables with normal and skewed distribution, respectively. Differences between lean and obese subjects were determined using the Wilcoxon test (a) and Student's T-test (b). Significant differences ($p < 0.05$) are marked in bold. Obesity criteria use BMI > 30 kg/m^2.

Microbiota of the Bali cohort

We produced over 8.5 Gb pair-ended reads of the 16S rRNA gene V4 region from 41 faecal metagenome extract, as previously described [26]. Quality and chimera filtering of these sequences yielded 6,655,198 pair-joined reads, between 110,358 and 228,531 reads per sample. Clustering at 97% sequence similarity yielded 3,167 non-singleton *de novo* Operational Taxonomic Units (OTUs). The OTUs were assigned to 13 Phyla, 55 Families, and 99 Genera according to greengenes reference data (version 3.18).

Using alpha diversity metrics as indicators of how community structure variation across the data set, Shannon entropy showed a range of values from 5 to 8 that were stable with sampling depth (Fig. 1a). Other metrics including richness (Chao1) and phylogenetic distance (Faith's PD) continued to increase as the sequencing depth increased (Additional File: Fig. S2). We normalized the data to 90,000 sequences per sample in further analyses.

Analyses of between-subject diversity revealed a robust separation of the microbiota samples into two groups based on community structure (Additional File1: Fig. S3). There was no microbiota distinction between individuals who are ethnically Balinese and otherwise. The separation into two clusters was strongly associated with relative abundances of the genera *Prevotella* and *Bacteroides* (Fig. 1b). The pattern is also observable in linkage analysis of relative abundances at the Family level (Fig. 2). We

consider these groups similar to the previously reported *Prevotella* and *Bacteroide*s enterotypes by Arumugam *et al.* [15]. Henceforth, we refer to these clusters of Bali microbiotas as Type-P and Type-B communities as recommended by Costea *et al.* [17]. There was a comparable number of Type-P (n = 18) and Type-B (n = 23) individuals. We confirmed that relative abundances of *Prevotella* and *Bacteroides* were significantly different ($p < 0.001$; Wilcoxon test) between groups categorized by community type (Additional File 1: Fig. S4).

To test for other community dimensions that may correlate with Type-P and Type-B individuals, we performed univariate vector regression analyses on genus-level abundance data (comprised of 80 most common OTUs) against the inter-sample weighted UniFrac distance matrix. The result identified eight OTUs as significant drivers of differences: *Bacteroides, Bilophila, Haemophilus*, an unclassified *Mogibacteriaceae* OTU, *Prevotella*, an unclassified *Prevotellaceae* OTU, an unclassified *Rikenellaceae* OTU, and *Sutterella* (all $p < 0.001$, Additional File 1: Table S1). To quantify the strength of the association in a multi-variable setting, we performed PERMANOVA (Additional File 1: Table S2). We found that compared to other OTUs, the association with *Prevotella* was the strongest ($R^2 = 0.33$, $p < 0.0001$), followed by *Bacteroides* ($R^2 = 0.16$, $p < 0.0001$). These results collectively suggest that although multiple dimensions of community structure are correlated with the categorization as Type-P or Type-B, *Prevotella* and *Bacteroides* correlate the strongest.

Individuals with Type-P communities had lower diversity

Diversity has widely been reported as a measure relevant to 'microbiota health'. Although many studies have found significant associations between microbiota diversity and health metrics, few patterns that are consistent across disparate human populations have emerged. The ecological significance of alpha diversity metrics is through their insight into the partitioning of ecological space between distinct taxa in a single community.

Here we found a consistent association between community type classification and alpha diversity metrics, whereby Type-P individuals were significantly lower in microbial diversity (Shannon) compared to lean Type-B individuals ($p < 0.001$, Wilcoxon test, Fig. 1c). Similar results were also seen using Faith's PD, Chao1, and Berger-Parker diversity metrics, albeit at lower statistical support (all $p < 0.05$, Wilcoxon test, Additional File 1: Fig. S5).

We also tested the relationship between obesity and diversity. Of the six obese individuals in our dataset, five were observed with Type-P microbiota and only one with Type-B (Fig. 1b). Given the effect of community type on diversity, this small number limits the ability to robustly test the comparison. Nonetheless, the Type-P individuals obese (n = 5) group had significantly lower Shannon diversity compared to the Type-P lean (n = 13) group ($p < 0.05$, Wilcoxon test). However, the difference was not statistically significant when measured using Faith's PD, Chao1, and Berger-Parker metrics ($p > 0.05$, Wilcoxon test).

Additionally, we observed a lower phylum-level richness in obese individuals regardless of community types (Additional File 1: Fig. S6). Yet, there was no significant difference in abundance for *Firmicutes, Bacteroidetes, Proteobacteria*, and *Actinobacteria* between the obese (n = 5) and lean (n = 35) group ($p > 0.05$, Wilcoxon test). Neither do we find a significant difference in the *Firmicutes* to *Bacteroidetes* ratio of these individuals ($p > 0.05$, Wilcoxon test). At the genus-level, 13 taxa were different by obesity ($p < 0.05$, Wilcoxon test), but these differences were not significant after adjustment for multiple comparisons (Additional File 1: Table S3).

Type-P and Type-B microbiotas may reflect distinct assembly processes

The Bali microbiotas were further explored for co-abundance relationships, which identified ten distinct genus-level co-abundance groups referred to as CAG1 through CAG10 (Fig. 3a, Additional File 1: Fig. S7). Both *Bacteroides* and *Prevotella* showed significant positive correlations (Kendall t > 0.5) to multiple other taxa forming CAG1 and CAG6, respectively. Using the net relative abundance of the CAGs (determined by summing the abundances of all CAG members in a sample), we observed that the CAGs containing *Prevotella* (CAG6) and *Bacteroides* (CAG1) had the highest summed abundances (Fig. 3a). As expected, their distribution across individuals reflected the Type-P and Type-B classification (Fig. 3b). Notably, these two CAGs and three others (CAG3, CAG4, and CAG7) were significantly different between the two community types (Fig. 3c). These five groups account for variations in 44 genus-level OTUs, although only 15 OTUs were independently different between Type-P and Type-B group after adjustment for multiple comparisons (Additional File 1: Table S4).

These results collectively suggest that the community type concept reflects fundamental differences in the outcomes of community assembly processes, and it is not simply a product of variance in the two most abundant genera (*Prevotella* and *Bacteroides*). Nevertheless, the method to consistently define community types is challenging; and we note that if CAG profiles were taken as the main criterion for community type classification (rather than the clustering approach used here), then there were 11 individuals in which neither CAG1 nor CAG6 was dominant (*i.e.* would not classify as Type P or B). For instance, three individuals we categorized as Type-P (BA013, BA028, and BA021) were outliers from the other Type-P's because of their lower CAG6 abundance (< 20%); and one individual we categorized as Type-P (BA019) is potentially a false-negative in that CAG6 had higher abundance than CAG1 (Fig. 3b). We also observed that in four Type-B individuals (BA002, BA006, BA015, and BA027) CAG4 was strongly present (Fig. 3b). The CAG4 is dominated by *Ruminococcus*, which has been proposed to represent a third enterotype [15, 17].

Through multivariate analyses, we found that CAG9 did not show significant associations with diet and community types (Additional File 1: Fig. 8); but it did show a significant association with obesity (Estimate = 0.52, $R^2 = 0.30$, $p < 0.05$). CAG9 association with increased BMI is retained in the all-inclusive model (Estimate = 0.49, $R^2 = 0.3$, $p < 0.05$). Across all variations of regression models, we only found an association between higher macronutrient intake with lower CAG1 abundance, but the relationships were

linked to community types and weaker in predictive value (Additional File 1: Fig. 8). In a univariate analysis, there was a trend for lower protein intake in Type-P individuals, but the effect was weak, and the relationship is not significant (Estimates= -2, R^2 = 0.85, p > 0.05). These results indicate that associations between diet and community types are negligible in our data, but associations between CAGs and physiological states (e.g. obesity) may exist.

Prevotella and *Bacteroides* gradient across populations

To look for similar patterns in community type distribution in different human populations, we conducted a meta-analysis with data from five other populations which have different ethnogeographic background. In addition to the 41 Bali, the data comprised of 22 Hadza hunter-gatherers, 21 Malawians, 28 Guahibo Amerindians, 16 Italians, and 129 Americans. The total number of samples is 257 individuals with 413,126,162 sequences – in which we observe 374 assigned taxonomic units (ATU) at species-level (Additional File 1: Table S5).

The result showed a separation of samples by demographic groups (Additional File 1: Fig. S9a). Interestingly, despite their narrow demographic structure (age, gender, and residency), the Bali microbiotas were a highly dispersed group in PCoA (Additional File 1: Fig. S9a). The pairwise inter-individual distance of the Bali was significantly higher than Malawi, Hadza, Italian, and USA (Additional File 1: Fig. S9b).

A known caveat for enterotypes is that they are defined with respect to the dataset that they are part of. Thus, categorizations in different studies are not directly comparable. To address this, we looked at the distribution of *Prevotella* and *Bacteroides* across all populations as a proxy indicator for Type-P and Type-B communities. We tested the applicability of this using Generalized Additive Modelling (GAM), in which the relative abundances of *Prevotella* and *Bacteroides* were mapped onto the PCoA data of weighted UniFrac distances (Fig. 4a-b). The first two principal coordinate axes account for 49.1% of the variance in the data (at 36.2% and 12.9% for PCo Axis 1 and 2, respectively). The results showed a strong fit for *Prevotella* and *Bacteroides* abundance to PCo Axis 1 and 2, respectively (adjusted R^2 > 0.9, p < 0.001, Additional File 1: Table S6-S7).

When the abundance distribution of *Prevotella* and *Bacteroides* was assessed by population, we found that communities with a dominance of *Prevotella* over *Bacteroides* were strongly over-represented in the three pre-transition societies (Hadza, Malawian, and Guahibo Amerindian) and the opposite for Americans and Italians (Fig. 4c). Consequently, the mean abundance of these genera was significantly different at a population level (Additional File 1: Fig. S10). The Bali cohort was distinctive in that *Prevotella* or *Bacteroides* dominance occurred in similar frequency in the population, and thus the mean abundances were similar.

Discussion

Nutrition-related chronic diseases are widely considered to be examples of dysbiosis in which their incidence is strongly shaped by socio-economic influences on the human diet. In this study, we aim to identify phenomena associated with nutrition transition that may drive microbiota patterns at the level of society. Particularly, we looked for microbiota features distinguishing pre-transition societies from those of post-transition; and further, to explore the dimensions of the environment which are most likely to have driven the difference. We postulated that features discriminating pre- and post-transition should show intermediate values in samples of societies actively undergoing transition. If true, then sampling such active-transition societies is predicted to provide a more useful dynamic response range for the identification of environmental drivers.

To test our hypotheses, we focussed on the concept of microbial enterotypes. It has been proposed that categorizing microbiotas into subtypes based on dominant features of their community structure (the enterotype concept) is a potential discriminator between pre- and post-transition societies [17]. Here, we tested this proposition, and its potential correlation to diet and disease, in a cohort of young female from Bali which represents a narrow demographic group from a society that has only very recently undergone transition [23, 24, 40, 41]. The analysis of diet and microbiota characteristics in this distinctive cohort, integrated with a meta-analysis of representative human microbiota datasets, provides new insights into potential drivers of enterotypes and their relevance to public health application.

The enterotypes concept was first proposed by Arumugam *et al* (2011) who coined the term to distinguish three clusters found in a multi-country human gut microbiota dataset based on Jensen-Shannon distance [15]. Similar patterns have since been reported in numerous other studies [5, 14, 16, 17]. However, there has been much controversy over how to use the concept of enterotypes as a tool to simplify the description of patterns in the human microbiota. It is evident that enterotypes are not intrinsically discrete biological entities; rather their observation is a property of the methods used to visualize beta diversity in the data [16, 17]. Consequently, a robust definition of enterotypes for the purpose of assessing their representation across populations in meta-analyses is very challenging. In a recent perspective article, Costea *et al.* proposed a unified nomenclature and guidelines for categorization aimed at facilitating the use of the concept for diverse datasets [17]. They proposed the terms ET-P, ET-B, and ET-F for referring to clusters where the attractor for that cluster is the taxon Genus:*Prevotella*, Genus:*Bacteroides*, or Class:*Firmicutes*. In this study, we have largely adhered to those guidelines, although taken a distinct approach to assessing the application of the concept in our meta-analysis using generalized additive models of marker taxa to support the findings.

In our beta diversity analyses of the Bali cohort, we found strong support for segregation into two clusters, which we refer to as Type-P and Type-B in accord with the proposed enterotype nomenclature. We used rigorous statistical tests to scrutinize these groupings, which showed that *Prevotella* and *Bacteroides* were indeed the main drivers of the clustering (although not the sole driver). These two genera have been consistently reported to differ between industrialized and subsistent societies [1, 5, 7, 11, 15–17, 42]. In our network analysis, co-abundance patterns were generally consistent with the cluster analysis – *Prevotella* or *Bacteroides* as the ‘hub genus’ were significantly over-represented in individuals

categorised to the designated community type. However, there were samples in which neither *Prevotella* nor *Bacteroides* are dominant. Although we did not see a third cluster in the Bali cohort, we note that four individuals in the Type-B cluster were dominated by CAG4, which includes various genera typically associated with the proposed third enterotype (ET-F). Our data are thus consistent with the existence of three similar types of community structure, but their visualization through clustering methods was not possible in the Bali dataset.

We then tested the proposition that the enterotypes would discriminate pre- and post-transition societies; and furthermore, that metrics based on these community types would be related to the stage of nutrition transition in that society. A challenge for this meta-analysis is that both the enterotype assignment and the determination of CAGs share a similar limitation in that they are products of a data-dependent clustering approach. As such, they are not suitable for a meta-analysis across multiple datasets. As a result, we did not see clusters that would permit facile classification into two or three community types in our analyses of the combined dataset, nor was this found in a similar study by Gorvitoskaia *et al.* [5]. As an alternative, we used the hub taxa, *Prevotella* and *Bacteroides*, to quantify Bali's position in the multi-country dataset and projected their relative abundance data onto the ordination using regression models. The result showed that *Prevotella* and *Bacteroides* abundances were strongly correlated to major differences in microbiotas across six populations.

Our data also showed that *Prevotella* and *Bacteroides* abundance in Bali represented an intermediate value if compared to pre- and post-transition societies. Our findings are consistent with our hypothesis that a nutritional transition is associated with a gradient of diverging communities states (measured here as enterotype and hub taxon incidence) – whereby the Bali cohort broadly occupied an intermediate position between pre- and post-transition societies in the multi-country dataset. These findings suggest that Bali is undergoing a shift in the meta-microbiota of their society. A plausible explanation for this is that recent industrialization of the food system have driven changes in microbial exposure and diet [23, 24, 40, 41], both of which are known to influence community assembly.

It has previously been found that enterotypes correlate to diet, including choices between animal or plant-based food [13, 14, 22]. Due to its association with nutrition, enterotypes has also been associated with increased incidence of cardiometabolic disorders (e.g. obesity, cholesterol) [17, 43–46]. To investigate the extent to which nutrition may alter the microbiota and affect obesity incidence, we employed various combinations of multivariate regression models on the Bali dataset. Our findings in these explorative analyses showed a striking lack of significant correlation between CAGs or community type to contemporary diet. Neither do we find associations with carbohydrate and fibre source (primarily rice). We found only a weak association between *Bacteroides* enrichment and higher protein intake (predominantly meat in the Bali dataset).

The lack of significant patterns may be due to the limited variety of food items in the Bali cohort, and thus these findings would require verification in a larger sample set. Nevertheless, our data do not support the idea that patterns of community type (or marker taxa) within a society will be reflected by short-term

diet observations. This does not exclude effects of long-term diet as a driver microbiota differences across populations. It is possible that different patterns of *Prevotella* and *Bacteroides* relative abundance between subsistence and industrialized societies may reflect different levels of fibre intake, animal-based food intake, and fat-sugar-enriched processed food products [5, 13, 14], but the drivers of this difference are likely to be more complex than short-term diet habits.

Although our Bali cohort is all young females, all of them were raised during the period of rapid lifestyle transition. Given that: 1) the timescale over which diet influences the observation of enterotype-like patterns, or the time point in developmental history when the patterns become fixed, is not yet clear; 2) enterotypes can become stable in early life; and 3) trans-generational microbe exposure influences microbiota diversity; it is possible that differences in their early development may have driven the striking divergence in microbiota pattern. Our data are consistent with the Thai US immigration study, which reported that living in an industrialised society induced a replacement of *Prevotella* with *Bacteroides* in the human gut and that it progressed across generations [8]. Recent studies in the Central African Republic (BaAka Pygmies and Bantu populations) and Nepal (Himalayan populations), also showed similar transitional states in the microbiota between rural and urban populations [6, 9]. Collectively, these data indicate that changes in human socio-economic conditions over time are reflected in the frequency distribution of *Prevotella* and *Bacteroides* across human societies.

Claims for microbiota association with nutrition-related disease began with obesity [44, 47]. Although the small size of the obese cohort in this study limits our ability to detect robust differences, our findings are instructive with regards to the use of simplifying metrics in health associations. Five of the six obese Bali individuals in our study were identified with a Type-P community which had lower microbial diversity. But despite this difference in diversity, it is unclear whether the difference is exclusively due to obesity or because the trait is linked to Type-P community. These findings present a contrast to the prevailing view that microbiotas commonly found in pre-industrial societies, typified by Type-P (or ET-P), have higher diversity and lower risk for NRCD [17, 43–46]. However, we point out that the apparent associations of diversity and obesity with Type-P here could be a product of the categorization into community types. Those Type-P individuals classified as obese in our study were distinguished by higher abundance of CAG9 and it was only this CAG that showed any statistical support in multi-variate models.

Costea *et al.* have proposed that one means of circumventing the issue in enterotype categorization is by enterotype assignment through relation to a reference dataset [17]. In our view, this approach has potential merit, but it does not address the main limitation of the method. The concept of enterotyping (at least as it has so far been promulgated) is essentially simplifying community structure to one dominant dimension. Therefore, its application to predicting the influence of the microbiota in individuals (such as disease risk or treatment outcome) will be dependent on the assumption that a significant fraction of relevant influence of the microbiota for that outcome is captured in the set of one-dimensional categories. Even in our small data set, it is clear that relying solely on enterotype classifications obscures the multi-dimensional nature of microbiota contribution to health outcomes [16]. However, our data also show that enterotype markers do have a strong predictive value for some population-level characteristics. We

propose that the concept of enterotypes has very limited value in predicting properties of an individual, but that it is useful in predicting properties of a demographic group and thus enterotypes may inform study design.

Together with other findings, our data highlighted the biological significance of enterotypes, particularly their implications toward disease association in future studies. Failure to address population heterogeneity between experimental groups may lead to false discoveries of disease biomarkers, particularly in societies undergoing socio-economic change. As shown in this study, the presence of two distinct community types in the Bali microbiota is a significant confounder for identifying obesity markers. We anticipate that microbiota disease association studies will be more robust if consideration is given to the baseline distribution of enterotypes (or proxy taxa) – since the enterotype is predicted to reflect extrinsic and intrinsic factors influencing community composition in the sampled population. In terms of human societies, these factors can include, but not limited to, ethnicity, cultural restrictions, early-life development, nutritional choices, other demographic factors, disease co-factors, pre-existing community composition, and inter-species relationships within the microbiota.

It has long been recognized that clinical trials require treatment groups to be matched for confounding factors such as age, gender, and socioeconomic status. An emerging model is that the gut microbiota exhibits the property of multi-stability; but in the stable states adopted, a large proportion of microbiota variance that occurs across human societies can be explained by the nutrient environment and inter-species interactions influencing the assembly. Whilst acknowledging the caveats described above, we propose that the enterotype concept can potentially underpin the development of useful tools to facilitate validation of life-history matched cohorts in studies of NRCD. This may be achievable through relatively simple proxy measures of enterotyping, such as abundance-ubiquity relationships of *Prevotella* in study cohorts; or simple indices for *Prevotella* and *Bacteroides* [5, 8]. Such microbiota-matched cohorts could inform the design of clinical trials and ultimately development of precision medicine strategies for obesity and comorbid diseases.

Conclusions

Bali is a society that has recently undergone transitions in their socioeconomic and nutritional landscape. We analysed the gut microbiota in 41 Bali individuals and found strong segregation into enterotype-like clusters that we refer to as Community Type-P (*Prevotella*-dominant) and Type-B (*Bacteroides*-dominant). Our meta-analysis incorporating data from 5 other distinct populations further showed that this segregation was strongly associated with differences in pre- and post-industrial populations, indicating that Bali is undergoing a shift in the meta-microbiota along with the changes in the society. Further analyses with diet and obesity data showed that the presence of two distinct community types in Bali is a significant confounder for identifying health markers, and thus enterotypes has limited value in predicting properties of an individual. Collectively, however, our findings suggest that microbiota clusters are much more useful in predicting population-level stratification that may inform demographic heterogeneity in future studies.

List Of Abbreviations

AMDR: Acceptable Macronutrient Distribution

ANOSIM: Analysis of Similarity

BMI: Body Mass Index

CAG: Co-Abundance Group

GAM: Generalized Additive Model

FDR: False Discover Rate

NRCD: Nutrition-Related Chronic Diseases

PCoA: Principal Coordinate Analysis

PERMANOVA: Permutational Multivariate Analysis of Variance

QIIME: Quantitative Insights Into Microbial Ecology

Declarations

Ethics

The ethical permit for this study was granted by the Udayana University Faculty of Medicine and Sanglah Hospital Ethics Commission on 18 September 2014 in Denpasar, Indonesia (No. 1286/UN.14.2/Litbang/2014). The permit was endorsed by the Eijkman Institute Research Ethics Commission on 24 December 2014 in Jakarta, Indonesia (Permit No. 80).

Consent for publication

This manuscript contains de-identified individual data. We have received written consent for publication of said data from all study participants.

Availability of data and materials

The Bali dataset supporting the conclusions of this article is available in the European Nucleotide Archive repository, project code PRJEB32385. Datasets for other populations are available in the MG-RAST repository, project number mgp401 and mgp7058. The authors declare that all data supporting the conclusions of this article are included within the paper and its additional files. Full accounts of the R scripts used for statistical analyses and data visualisation is available in Additional File 2.

Competing Interest

All authors declared no competing interest.

Authors' Contributions

CAF, SGM, AJH, IWW, DMW, and HS performed sampling. CAF proposed the study, did laboratory work, analysed the data, and drafted the manuscript. SGM, HS, and AJH designed, directed, and facilitated the study, and provided major support in the data interpretation. DMW and IWW coordinated and facilitated the study enrolment, including the collection and interpretation of diet and demographic data. RD, and RM provided support for the analysis and interpretation of diet data. All authors reviewed and approved the final version of the manuscript.

Funding

This research was partially supported by the Indonesia Ministry of Research and Technology / National Agency for Research and Innovation through the Eijkman Institute for Molecular Biology; Australia's Department of Foreign Affairs and Trade through the Australia Awards Scholarship; and the University of Sydney International Program Development Fund (2013 Round).

Acknowledgments

We are grateful to university students at the Faculty of Medicine, Udayana University, Denpasar (Tjokorda Istri Pramitasuri, and colleagues), who had assisted the study during the recruitment, enrolment, and data collection stage in Denpasar. We thank Prodia Laboratory in Denpasar for supporting the collection and initial storage of the samples. We are grateful to Prof. Sangkot Marzuki at Indonesia Science Academy (Akademi Ilmu Pengetahuan Indonesia) for facilitating and supporting to the commencement of this study. We thank our colleagues, Sukma Oktavianthi, Lidwina Priliani, Hidayat Trimarsanto, Eline Klaassens, and Mark Read for their input and guidance.

References

1. Filippo CD, Cavalieri D, Paola MD, Ramazzotti M, Poullet JB, Massart S, et al. Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. PNAS. 2010;:201005963.
2. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, et al. Human gut microbiome viewed across age and geography. Nature. 2012;486:222–7.
3. Schnorr SL, Candela M, Rampelli S, Centanni M, Consolandi C, Basaglia G, et al. Gut microbiome of the Hadza hunter-gatherers. Nat Commun. 2014;5:3654.

4. Martínez I, Stegen JC, Maldonado-Gómez MX, Eren AM, Siba PM, Greenhill AR, et al. The gut microbiota of rural papua new guineans: composition, diversity patterns, and ecological processes. Cell Rep. 2015;11:527–38.
5. Gorvitovskaia A, Holmes SP, Huse SM. Interpreting Prevotella and Bacteroides as biomarkers of diet and lifestyle. Microbiome. 2016;4:15.
6. Gomez A, Petrzelkova KJ, Burns MB, Yeoman CJ, Amato KR, Vlckova K, et al. Gut Microbiome of Coexisting BaAka Pygmies and Bantu Reflects Gradients of Traditional Subsistence Patterns. Cell Reports. 2016;14:2142–53.
7. Mancabelli L, Milani C, Lugli GA, Turroni F, Ferrario C, van Sinderen D, et al. Meta-analysis of the human gut microbiome from urbanized and pre-agricultural populations. Environ Microbiol. 2017;19:1379–90.
8. Vangay P, Johnson AJ, Ward TL, Al-Ghalith GA, Shields-Cutler RR, Hillmann BM, et al. US Immigration Westernizes the Human Gut Microbiome. Cell. 2018;175:962-972.e10.
9. Jha AR, Davenport ER, Gautam Y, Bhandari D, Tandukar S, Ng KM, et al. Gut microbiome transition across a lifestyle gradient in Himalaya. PLOS Biology. 2018;16:e2005396.
10. Popkin BM. Nutrition Transition and the Global Diabetes Epidemic. Curr Diab Rep. 2015;15:64.
11. Obregon-Tito AJ, Tito RY, Metcalf J, Sankaranarayanan K, Clemente JC, Ursell LK, et al. Subsistence strategies in traditional societies distinguish gut microbiomes. Nat Commun. 2015;6:1–9.
12. Human Microbiome Project Consortium T. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486:207–14.
13. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, et al. Diet rapidly and reproducibly alters the human gut microbiome. Nature. 2014;505:559–63.
14. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen Y-Y, Keilbaugh SA, et al. Linking long-term dietary patterns with gut microbial enterotypes. Science. 2011;334:105–8.
15. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, et al. Enterotypes of the human gut microbiome. Nature. 2011;473:174–80.
16. Knights D, Ward TL, McKinlay CE, Miller H, Gonzalez A, McDonald D, et al. Rethinking “enterotypes.” Cell Host Microbe. 2014;16:433–7.
17. Costea PI, Hildebrand F, Arumugam M, Bäckhed F, Blaser MJ, Bushman FD, et al. Enterotypes in the landscape of gut microbial community composition. Nat Microbiol. 2018;3:8–16.
18. Sonnenburg ED, Sonnenburg JL. Starving our Microbial Self: The Deleterious Consequences of a Diet Deficient in Microbiota-Accessible Carbohydrates. Cell Metab. 2014;20:779–86.
19. Sonnenburg ED, Smits SA, Tikhonov M, Higginbottom SK, Wingreen NS, Sonnenburg JL. Diet-induced extinctions in the gut microbiota compound over generations. Nature. 2016;529:212–5.
20. Roager HM, Licht TR, Poulsen SK, Larsen TM, Bahl MI. Microbial Enterotypes, Inferred by the Prevotella-to-Bacteroides Ratio, Remained Stable during a 6-Month Randomized Controlled Diet Intervention with the New Nordic Diet. Appl Environ Microbiol. 2014;80:1142–9.

21. Rajilić-Stojanović M, Heilig HGHJ, Tims S, Zoetendal EG, Vos WM de. Long-term monitoring of the human intestinal microbiota composition. Environmental Microbiology. 2013;15:1146–59.
22. Wang J, Linnenbrink M, Künzel S, Fernandes R, Nadeau M-J, Rosenstiel P, et al. Dietary history contributes to enterotype-like clustering and functional metagenomic content in the intestinal microbiome of wild mice. Proc Natl Acad Sci USA. 2014;111:E2703-2710.
23. Koninck RD, Déry S. Agricultural Expansion as a Tool of Population Redistribution in Southeast Asia. J Southeast Asian Stud. 1997;28:1–26.
24. Antara M, Sumarniasih MS. Role of Tourism in Economy of Bali and Indonesia. Journal of Tourism and Hospitality Management. 2017;5:33–44.
25. Juergen Erhardt. NutriSurvey: Nutrition Surveys and Calculations [Computer software]. English. Germany: EBISpro; 2010. http://www.nutrisurvey.de/index.html.
26. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, et al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J. 2012;6:1621–4.
27. Aronesty E. ea-utils: “Command-line tools for processing biological sequencing data.” 2011. https://code.google.com/p/ea-utils/. Accessed 11 Jun 2020.
28. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Meth. 2010;7:335–6.
29. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics. 2011;27:2194–200.
30. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. Appl Environ Microbiol. 2006;72:5069–72.
31. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. Appl Environ Microbiol. 2007;73:5261–7.
32. Faith DP, Baker AM. Phylogenetic diversity (PD) and biodiversity conservation: some bioinformatics challenges. Evol Bioinform Online. 2007;2:121–8.
33. Lozupone C, Lladser ME, Knights D, Stombaugh J, Knight R. UniFrac: an effective distance metric for microbial community comparison. ISME J. 2011;5:169–72.
34. Kindt R, Coe R. Tree diversity analysis: A manual and software for common statistical methods for ecological and biodiversity studies. Nairobi, Kenya: World Agroforestry Centre; 2005.
35. Murtagh F, Legendre P. Ward’s Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward’s Criterion? J Classif. 2014;31:274–95.
36. Wilke A, Bischof J, Gerlach W, Glass E, Harrison T, Keegan KP, et al. The MG-RAST metagenomics database and portal in 2015. Nucleic Acids Res. 2016;44:D590-594.
37. Benjamini Y, Yekutieli D. The Control of the False Discovery Rate in Multiple Testing under Dependency. Ann Statist. 2001;29:1165–88.

38. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, et al. vegan: Community Ecology Package. R package version 2.4-3. English. Finland; 2017. https://CRAN.R-project.org/package=vegan. Accessed 11 Jun 2020.
39. Kloke JD, McKean JW. Rfit: Rank-based Estimation for Linear Models. The R Journal. 2012;4:8.
40. Statistics Indonesia. Number and Growth Rate of Populations in Denpasar Municipality,2001-2015. https://denpasarkota.bps.go.id/statictable/2016/07/25/157/jumlah-dan-laju-pertumbuhan-penduduk-kota-denpasar-2001-2015.html. Accessed 11 Jun 2020.
41. Statistics Indonesia. Regional GDP of Denpasar 2010-2018. https://denpasarkota.bps.go.id/dynamictable/2019/07/29/86/pdrb-kota-denpasar-atas-dasar-harga-berlaku-menurut-lapangan-usaha-tahun-2010-2018-juta-rupiah-.html. Accessed 11 Jun 2020.
42. Koren O, Knights D, Gonzalez A, Waldron L, Segata N, Knight R, et al. A Guide to Enterotypes across the Human Body: Meta-Analysis of Microbial Community Structures in Human Microbiome Datasets. PLoS Comput Biol. 2013;9:e1002863.
43. Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, et al. Richness of human gut microbiome correlates with metabolic markers. Nature. 2013;500:541–6.
44. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, et al. A core gut microbiome in obese and lean twins. Nature. 2009;457:480–4.
45. Wang J, Li W, Wang C, Wang L, He T, Hu H, et al. Enterotype Bacteroides Is Associated with a High Risk in Patients with Diabetes: A Pilot Study. J Diabetes Res. 2020;2020.
46. de Moraes ACF, Fernandes GR, da Silva IT, Almeida-Pititto B, Gomes EP, Pereira A da C, et al. Enterotype May Drive the Dietary-Associated Cardiometabolic Risk Factors. Front Cell Infect Microbiol. 2017;7. doi:10.3389/fcimb.2017.00047.
47. Ley RE, Bäckhed F, Turnbaugh P, Lozupone CA, Knight RD, Gordon JI. Obesity alters gut microbial ecology. PNAS. 2005;102:11070–5.

Figures

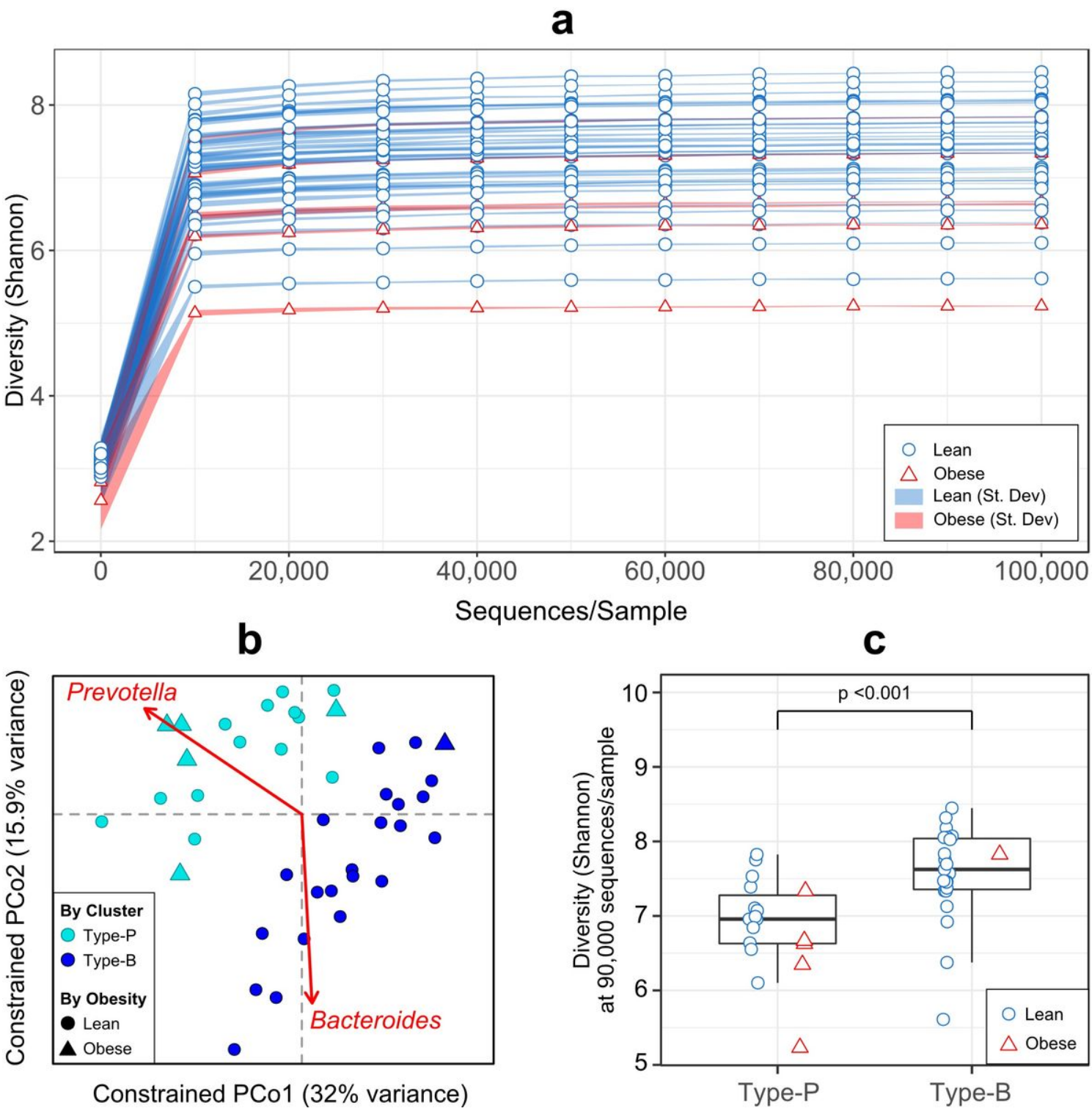

Figure 1

Diversity of the gut microbiota in 41 Bali individuals. a, Shannon diversity at various sequencing depths per sample. Data points and line thickness reflects the mean and standard deviation (St. Dev). b, Constrained Principal Coordinate Analysis (PCoA) of weighted Unifrac distance at depth 90,000 sequences per sample. The constrains used Prevotella and Bacteroides abundance. Samples clustered into Type-P (n=18) and Type-B (n=23) groups based on Ward's clustering method. c, Comparison of Shannon diversity between Type-P and Type-B using the Wilcoxon test.

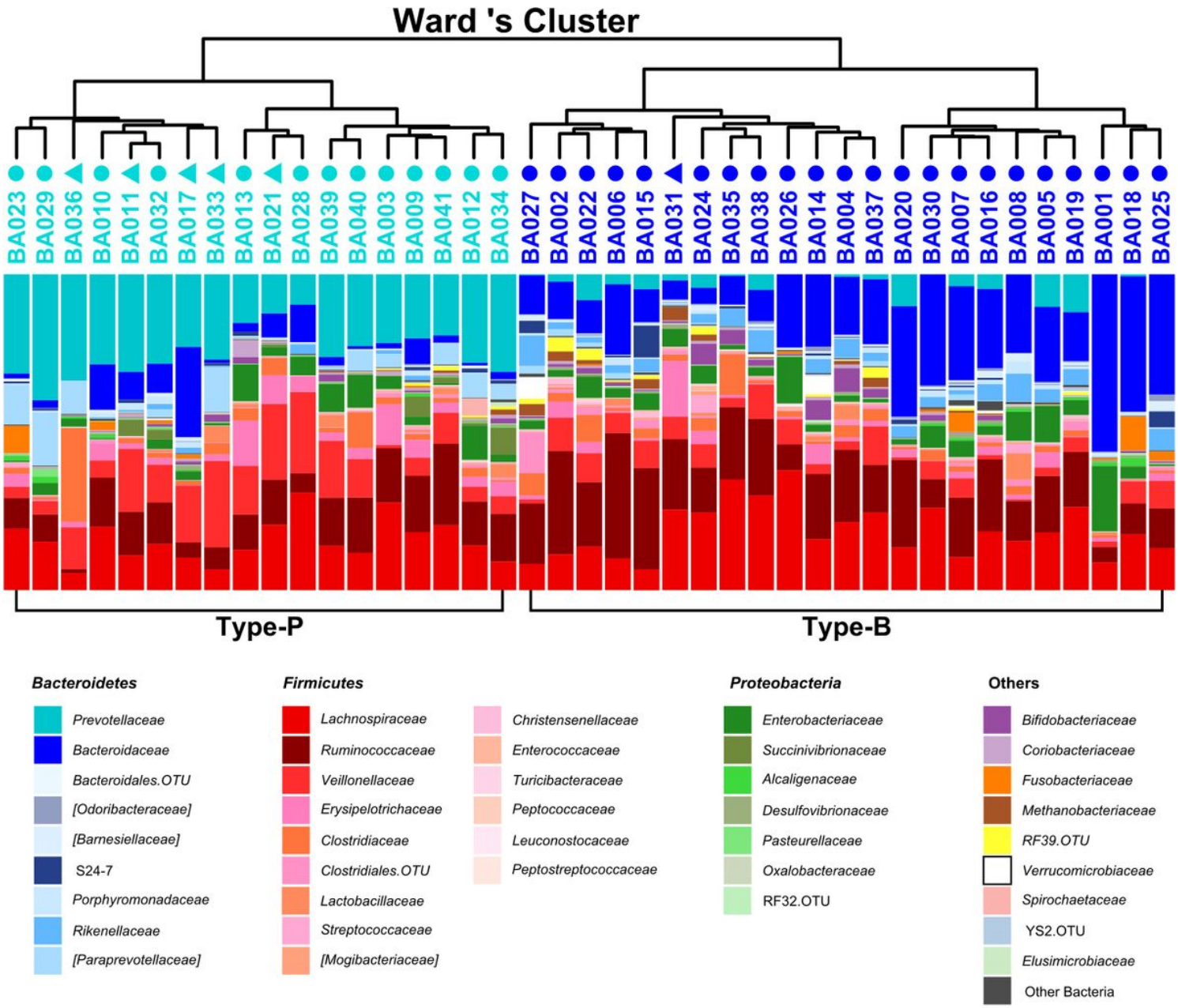

Figure 2

The relative abundance of bacterial families in 41 Bali individuals. The samples were ordered according to Ward's clustering method based on Weighted UniFrac data.

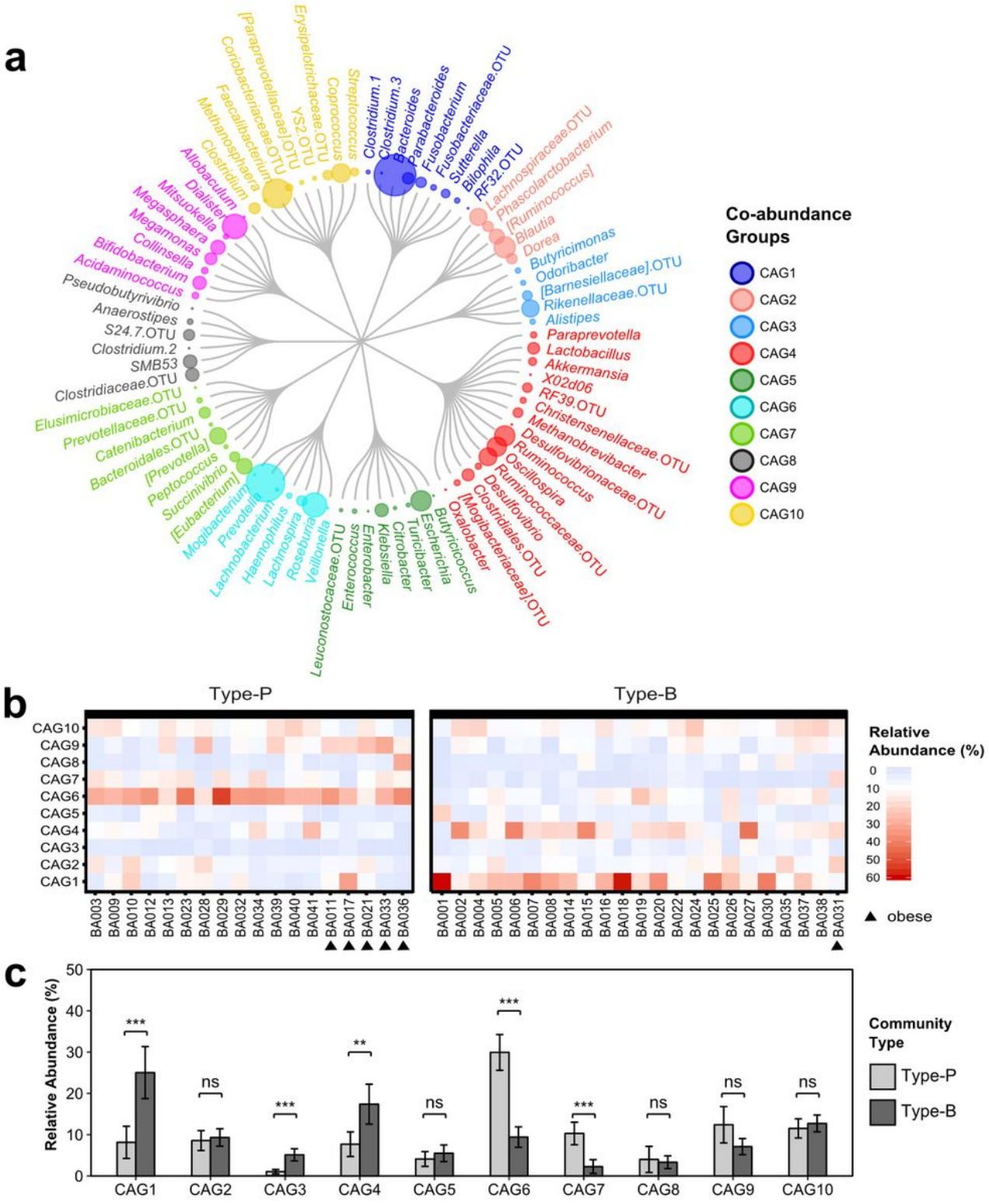

Figure 3

Genus abundance in 41 Bali individuals grouped by co-abundance pattern. a, A dendrogram of ten distinct co-abundance groups (CAG1 – CAG10). Each node represents a taxon, of which the size is proportional to abundance. b, Net abundance and distribution of CAG1 – CAG10 by community types. Box colour represents abundance (see legend). c, Comparison of CAG abundance in Type-P and Type-B

microbiota communities; error bars represent standard errors of the mean. Significant differences (Wilcoxon test) are marked with codes: ***, p <0.001; **, p <0.01; and ns, not significant.

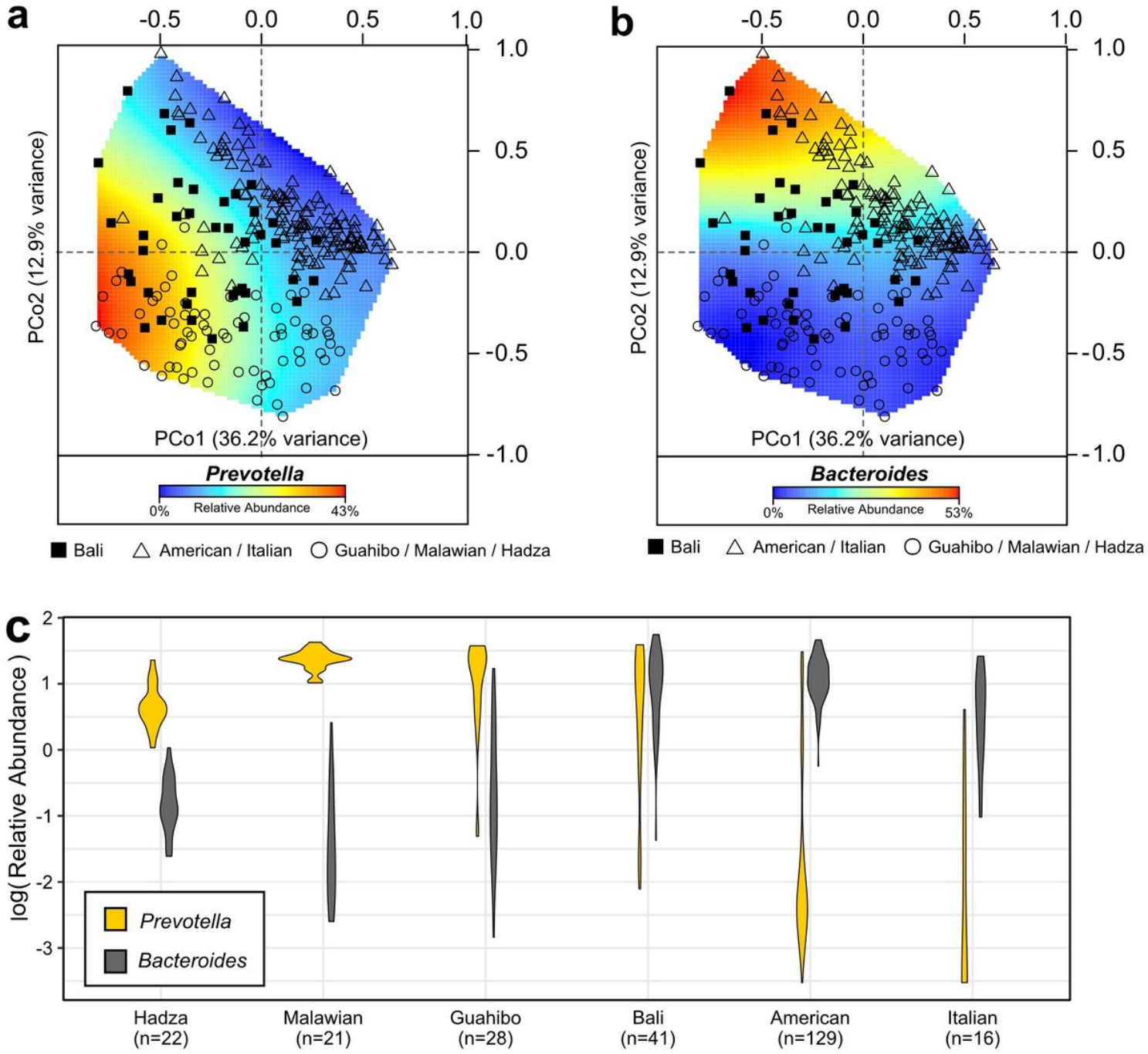

Figure 4

Comparison of Prevotella and Bacteroides abundance in 41 Bali individuals and 216 others, including 22 Hadza hunter-gatherers, 21 Malawians, 28 Guahibo Amerindians, 129 Americans, and 16 Italians. a, Prevotella abundance fitted into PCoA of weighted UniFrac data with Generalized Additive Model. The gradient of colours represents abundance (see legend). b, as in a, but the PCoA was fitted with Bacteroides abundance. c, Density profile of Prevotella and Bacteroides by samples and population of

origin. Wider sections of the violin plot indicate a higher number of observations at a given abundance value (log transformed), the thinner sections correspond to a lower number of observations.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- additionalfile8metadataallpopulation.xls
- additionalfile8metadataallpopulation.xls
- additionalfile7sixpopulationgenusrelativeabundance.csv
- additionalfile7sixpopulationgenusrelativeabundance.csv
- additionalfile6baliL6genusrelativeabundance.csv
- additionalfile6baliL6genusrelativeabundance.csv
- additionalfile5baliL2phylumabsabundance.csv
- additionalfile5baliL2phylumabsabundance.csv
- additionalfile4weightedunifracdata.xls
- additionalfile4weightedunifracdata.xls
- additionalfile3alpharefaction.xls
- additionalfile3alpharefaction.xls
- additionalfile2rscripts.pdf
- additionalfile2rscripts.pdf
- additionalfile1supplementaryresults.doc
- additionalfile1supplementaryresults.doc