Title

Oral microbiomes from hunter-gatherers and traditional farmers reveal shifts in commensal balance and pathogen load linked to diet

Florent Lassalle[1,2]*, Matteo Spagnoletti[1]*, Matteo Fumagalli[1], Liam Shaw[1], Mark Dyble[3,4], Catherine Walker[1], Mark G. Thomas[1], Andrea Bamberg Migliano[3], Francois Balloux[1].

[1]*University College London, UCL Genetics Institute, Gower Street, London WC1E 6BT, UK;*

[2]*Imperial College London, Department of Infectious Disease Epidemiology, Praed Street, London W2 1NY, UK;*

[3]*University College London, Department of Anthropology, 14 Taviton Street, London WC1H 0BW, UK;*

[4]*Institute for Advanced Study in Toulouse, 21 allée de Brienne, 31015 Toulouse Cedex 6, France.*

**These authors contributed equally to this work*

Keywords: metagenomics; hunter-gatherers; oral microbiome; diet; Philippines

Corresponding authors:

Florent Lassalle (f.lassalle@imperial.ac.uk); Francois Balloux (f.balloux@ucl.ac.uk)

Running title

Impact of diet on oral microbiome composition

Abstract

Maladaptation to modern diets has been implicated in several chronic disorders. Given the higher prevalence of disease such as dental caries and chronic gum diseases in industrialized societies, we sought to investigate the impact of different subsistence strategies on oral health and physiology, as documented by the oral microbiome. To control for confounding variables such as environment and host genetics, we sampled saliva from three pairs of populations of hunter-gatherers and traditional farmers living in close proximity in the Philippines. Deep shotgun sequencing of salivary DNA generated high-coverage microbiomes along with human genomes. Comparing these microbiomes with publicly available data from individuals living on a Western diet revealed that abundance ratios of core species were significantly correlated with subsistence strategy, with hunter-gatherers and Westerners occupying either end of a gradient of *Neisseria* against *Haemophilus*, and traditional farmers falling in between. Species found preferentially in hunter-gatherers included microbes often considered as oral pathogens, despite their hosts' apparent good oral health. Discriminant analysis of gene functions revealed vitamin B5 autotrophy and urease-mediated pH regulation as candidate adaptations of the microbiome to the hunter-gatherer and Western diets, respectively. These results suggest that major transitions in diet selected for different communities of commensals and likely played a role in the emergence of modern oral pathogens.

Introduction

Humans have experienced dramatic changes in diet over the last 10,000 years (Mathieson et al., 2015; Quercia et al., 2014). The Neolithic transition marked the beginning of wide-scale dietary and demographic changes from subsistence by primarily nomadic hunting and gathering to sedentary agriculture (Bocquet-Appel, 2011). A second, equally dramatic nutritional shift occurred with the Industrial Revolution in the mid-19th century, which led to widespread availability of processed flour and sugar (Cordain et al., 2005). These alterations of ancestral diets have been implicated in the emergence of modern chronic disorders, including cardiovascular disease, diabetes, obesity and osteoporosis (Cordain et al., 2005).

The human microbiome, the sum of diverse microbial ecosystems colonizing the various niches offered by the human body, is known to play an important role in human health (Lloyd-Price, Abu-Ali, & Huttenhower, 2016; Yang et al., 2012). In particular, the oral cavity, which is the gateway to the human body for both food and air intake, hosts the oral microbiome (Dewhirst et al., 2010). Shifts in composition of this microbial community have been associated with several oral conditions such as periodontitis (Griffen et al., 2012), which in turn is suspected as a cause of a series of modern chronic disorders, including inflammatory bowel disease, diabetes, cardiovascular disease and some forms of cancer (Kuo, Polson, & Kang, 2008; Li, Kolltveit, Tronstad, & Olsen, 2000; Whitmore & Lamont, 2014). By occupying a major interface between the human body and the external environment, the oral microbiome is shaped both by host variables, such as genetic background, general health and immunity, and by external environmental factors including ecology and diet. The relative abundance of microbes colonizing the mouth changes along the day through growth and regular clearance by swallowing of saliva, but the set of taxa observed over time in an individual's mouth is remarkably stable (Carpenter, 2013; Marsh, Do, Beighton, & Devine, 2016).

Despite its compositional stability on the shorter term, there is strong evidence that oral microbiome composition has been shaped by major sociocultural changes over our recent evolutionary history (Mira *et al.*, 2006; Hunter, 2014). Indeed, analysis of ancient and historic dental calculus samples has identified major shifts in species composition in the oral microbiome coinciding with the Neolithic and Industrial Revolution (Adler et al., 2013; Warinner et al., 2014). As dietary and oral hygiene standard shifts have occurred over a relatively short evolutionary timescale, it has been suggested that modern human microbiomes may be maladapted to current conditions, leading to increased incidence of oral diseases. This would be consistent with the spread of major oral polymicrobial diseases across human populations in recent times (Marsh, 2003; Zaura, Nicu, Krom, & Keijser, 2014). In most industrialized countries 60-90% of children have signs of caries and clinically defined periodontal disease is highly prevalent among adults (Petersen, 2005). Additionally, modern chronic disorders linked to oral disease – inflammatory bowel disease, cardiovascular disease, diabetes and cancer (Kuo, Polson, & Kang, 2008; Whitmore & Lamont, 2014) – all tend to be rare in contemporary hunter-gatherers, whose lifestyle and diet is deemed close to that of ancestral humans (Cordain et al., 2005). This suggests that the microbiome could act as a coupling link between human lifestyle and health. In particular, we make the hypothesis that recent changes in lifestyle and diet could have impacted the composition of oral microbiomes, which became conducive of modern chronic disorders.

The differences observed between archaeological and modern microbiomes may however not necessarily arise from shifts in subsistence strategies but from many other factors that changed through time. Additionally, direct comparison between modern microbiomes to those generated from ancient, degraded remains is not straightforward. It thus appears that comparison of microbiomes from contemporary populations exposed to similar environments but with contrasted lifestyles may represent the best experimental design to test whether diet is directly shaping the salivary microbiome. A series of studies have investigated the microbiome composition of modern hunter-gatherers in comparison to neighboring populations of traditional farmers or more distant Western individuals (Clemente et al., 2015; Morton et al., 2015; Nasidze et al., 2011; Obregon-Tito et al., 2015; Schnorr et al., 2014). Notably, a few studies detected an effect of subsistence strategy on the oral microbiome, highlighting composition trends, such as the increased abundance of Fusobacteriaceae, Prevotellaceae, *Veillonella* spp. and *Haemophilus* spp. in hunter-gatherers' oral microbiomes (Clemente et al., 2015; Nasidze et al., 2011). However, these common composition features may be largely coincidental and need to be compared with data from other settings to consider them as diagnostic of subsistence strategy.

In addition, comparisons of microbiomes for a single pair of populations (e.g. hunter-gatherer against a population having adopted a Western diet) are likely to be confounded by additional differences in geographical origin, health, socio-economic status and possibly genetic backgrounds between the populations. To circumvent these problems, we designed our study around three pairs of populations living in close proximity in the Philippines and sharing essentially the same environment: Batak and Tagbanua, Aeta and Zambal and Agta and Casigurani, respectively hunter-gatherers (HGs) and traditional farmers (TFs). This design allowed us to detect systematic differences between all three pairs of populations that are much more likely to be driven by subsistence strategy. We also relied on deep whole genome shotgun (WGS) sequencing rather than the more standard but limited 16S rRNA amplicon-sequencing (Clemente et al., 2015; Nasidze et al., 2011). While the additional costs of the shotgun sequencing limited study sample size, it comes with an increased ability to resolve microbial species composition – in particular for populations whose microbiomes have not been well characterized to date – and also opens the door to direct investigation of the biological functions involved in their adaptation. This WGS approach also allowed us to generate human genomes (2-20x depth), which we used to control for a possible effect of the host genetic make-up.

The high-coverage oral microbiomes we generated were combined with previous datasets obtained with a similar protocol from individuals from the USA subsisting on a Western diet (Hasan et al., 2014; The Human Microbiome Project Consortium, 2012). These data were processed with state-of-the-art taxonomic assignation and phylogenetic diversity analyses to tease apart the effect of diet, environment and human genetic make-up in shaping the composition of the oral microbiome.

Materials and Methods

Study design, subject enrollment and DNA collection

The study included 24 samples selected from a large collection of saliva samples (>350) collected during a long-term fieldwork project in the Philippines under the supervision of Dr. Andrea Migliano (Hunter-Gatherers' resilience Project). Aeta live in the mountain forests from the western part of Luzon island, and Agta are from the East of Luzon, close to the coast. Batak live in the mountain forests in the central part of the Palawan island; the TF groups (Zambal, Casigurani and Tagbanua) live in close geographical proximity (1-10km) to each of the respective neighboring HG groups (Fig. 1). Saliva samples were collected in 2007, 2008 and 2009 (Table S1). Using the Oragene DNA OG-500 collection kit (DNA Genotek, Kanata, Canada), participants were asked to wash their mouth with water and then to spit into the vial until it is half full. All the samples were transported to London UK, where they were stored at the UCL department of Anthropology at -20°C.

The protocol was in accordance with the Helsinki Declaration, and was approved by the Ethics Commission of the University College London, London UK. We further obtained ethical clearance from the National Commission on Indigenous Peoples (NCIP) (Cariño, 2012). Approval was also obtained at the local community level, from the elders' committee in each of the locations, and informed consent was obtained from all participants (written in their own languages) after a presentation of the research objectives in Tagalog for the Philippine populations; a copy of the Participant Information Sheet and an English version of the Participant Consent Form (blank copy) are available in Sup. File S1.

Sample selection and DNA extraction.

We selected four samples (from individual aged between 20 and 40 and in good oral health) for each group of hunter-gatherers and their neighboring farmers, generating three geographical groups of eight samples each, for a total of 24 samples. We randomly selected two males and two females, under the constraint of individuals being unrelated (without known family relationships, based on using the anthropological information collected during the field work). All the samples have been anonymized. DNA was purified from saliva employing the Oragene DNA isolation kit (DNA Genotek, Kanata, Canada), following the manufacturer's recommended instructions. DNA quantification and quality controls were accomplished using Qubit 2.0 fluorimeter (Thermo Fisher Scientific, Waltham, USA) and Agilent 2100 Bioanalyzer DNA chips (Agilent technologies, Santa Clara, USA).

DNA library preparation, sequencing and quality control

Aliquots of 1µg DNA per sample were used to create sequencing libraries. First, genomic DNA was fragmented using a Covaris S2 sonicator (Covaris Inc., Woburn, USA) to approximately 300bp. Fragmented DNA was quantified and used to synthesize shotgun libraries with the NebNext Ultra DNA library preparation kit for Illumina (New England Biolabs, Ipswich, USA), according to manufacturer's instructions. PCR cycling conditions were set to a minimum of 4 cycles for annealing/extension to minimize PCR duplicates. NEBNext Singleplex Oligos for Illumina were used for indexing samples without multiplexing. All the samples have been sequenced at the UCL Institute of Neurology using 100bp paired-end chemistry and the Illumina HiSeq 2500 system (Illumina, San

Diego, USA). Three libraries were prepared, each grouping eight individuals: library #1 (4 Aeta HGs + 4 Zambal neighboring TFs), library #2 (4 Batak HGs + 4 Tagbanua neighboring TFs) and library #3 (4 Agta HGs + 4 Casigurani neighboring TFs).
The libraries #1 and #2 were sequenced on one Illumina flow-cell each (8 lanes, one per individual), while the library #3 was sequenced in two rounds, using two flow-cells (16 lanes, two per individual). The whole sequencing process produced 21,362,688,072 reads (>870GB of data) passing filters (Illumina CASAVA 1.8.0, default settings). Raw reads were processed using the first step of the MOCAT pipeline (version 1.3) (Kultima et al., 2012) with standard settings (options "-identity 97 -length 45 -soapmaxmm 5"): reads were quality trimmed, adapters were removed, and so were reads matching human when mapping to reference hg19 (Genome Reference Consortium Human Reference [GRCh] 37) using SOAPAligner2 (Li et al., 2009) version 2.21 with options "-r 2 -M 4 -l 30 -v 5 -p 4". This reduced the dataset to a total of 1.13 billion reads, with 8.3—147.7 million reads per individual. These read sets were submitted to the ENA (www.ebi.ac.uk/ena) under the BioSample accessions ERS1202862—ERS1202885. Human-mapped reads were further used to analyze the genetic diversity of the sampled individuals (see Supplementary Methods).

Kraken reference database

We built a custom Kraken database (Wood & Salzberg, 2014) made from all available RefSeq genomes for bacteria (94,803), archaea (676), viruses (7,497), protozoa (79) and fungi (238) using the ncbi-genome-download application (https://github.com/kblin/ncbi-genome-download), as well as all available RefSeq plasmids (10,842) directly from the NCBI FTP server (ftp://ftp.ncbi.nih.gov/refseq/release/plasmid) as of September 19th 2017. We added the GRCh38, HuRef and YanHuang human genome reference sequences (International Human Genome Sequencing Consortium, 2004; Levy et al., 2007; Wang et al., 2008). The database was indexed for the distribution of 31-mers in reference genomes, using 15-bp minimizers (Wood & Salzberg, 2014). The full database had a final size of 539 Gb; this was shrunk to a 'Mini-kraken' indexed database of 193 Gb, covering 38,190 different taxa (with distinct NCBI taxon id).

Estimation of microbial taxonomic abundances

The 24 metagenomes generated in this study and nine additional Western metagenomes from other studies (see 'Public microbiome data' section in Supplementary text) were analyzed as follows. Reads were classified in terms of taxonomic origin using Kraken (Wood & Salzberg, 2014) version 0.10.6. This software searches k-mers in sequencing reads that match a custom database of reference genomes. Inclusion of the human genome in the reference database allowed to screen for remaining reads that were not identified at the previous filtering step by mapping. Reads assigned to human were removed from later steps of the analyses using a custom Python script (http://github.com/flass/microbiomes/kraken/parseKronaGetReadsByTaxid.py, option '--exclude.taxa 9606').

Kraken assigns reads to all taxonomic levels in a cumulative manner, and relative abundance of taxa can be computed using the ratio of read counts at one specific level over the total. Read counts were computed 1) with a conservative filter on read confidence scores, i.e. keeping only reads with more than 20% k-mers assigned to congruent taxa (using kraken-filter executable with option "--thresh

0.20"); and 2) in a sensitive mode, i.e. without confidence score filtering. Relative abundances were computed at the species and genus level. Distribution of relative species abundances per sample (from sensitive mode) showed significant bias relative to sequencing depth for values under 10^{-12}, with low-depth samples being depleted in rare species (Fig. S1), so the dataset was truncated to species relative abundance values above 10^{-12}, decreasing the number of represented species from 8,226 to 5,323. We used linear discriminant analysis (LDA) effect size (LEfSe) (Segata et al., 2011) to detect taxa that significantly differentiate groups of samples based on their subsistence strategy (accounting for the underlying grouping by population). We then used a simple LDA, as implemented in the ade4 R package (Dufour & Dray, 2007), to identify the species that specifically differentiate microbiomes along the human lifestyle gradient opposing HGs to Western controls (WCs); significance was assessed with pairwise t-tests, Wilcoxon rank-sum tests (using Benjamini-Hochberg false discovery rate [FDR] correction procedure for multiple testing) and ANCOM test (Mandal et al., 2015), with low stringency multiple testing correction (option 'multcorr=2'). Abundance tables and complete reports of statistical analyses for filtered and unfiltered dataset, at species and genus levels, are available on Figshare at: https://figshare.com/s/72d9a99703c222f3ecfb. Kraken taxonomic assignation makes use of the entire WGS dataset, allowing to characterize the presence of low-abundance organisms, but is biased towards taxa closely related to organisms represented in the reference database where exact sequence matches are possible, and does therefore not account for the phylogenetic sampling bias in the database.

We thus used Phylosift (Darling et al., 2014) (version 1.0.1) to characterize relative abundances of lineages in a phylogenetic placement framework that naturally allows for a robust assignment of taxonomic identity to sequences that are highly divergent with respect to the reference database. Briefly, a database of 33 highly conserved marker genes (Phylosift default built-in database, version 1395376975, available at http://edhar.genomecenter.ucdavis.edu/~koadman/phylosift_markers/) was searched for similarity with all reads. Those reads that matched (roughly 0.5-1% of the total dataset) were then assigned to a branch of a species tree built from the concatenation of the marker genes' reference alignments, using a phylogenetic placement algorithm (Matsen, Kodner, & Armbrust, 2010).

This procedure yielded a table of the density of placed reads per branch of the reference species tree, which can be used to compute relative abundances of a clade by summing the placement densities of all branches of the corresponding subtree. These can be translated into robust relative abundance estimates of named organisms at any taxonomic level using the taxonomic labelling of the branches of the tree provided with the Phylosift package, typically with a resolution of 10^{-3} for frequencies of named species. The structure of the diversity of microbiome composition among samples can be conveniently explored using principal component analysis (PCA) of the difference of placement densities between reference tree edges, hereafter referred as 'edge PCA'. Custom Python and R scripts using the ade4 package (Dufour & Dray, 2007) (http://github.com/flass/microbiomes/tree/master/scripts/phylosift) were used to select representative eigenvectors in the edge PCA (corresponding to branches of the reference tree) for graphical representation in a 2-D plane: among the set of all eigenvectors directed in the same quadrant of the plane that correspond to branches of a same clade in the species tree, the eigenvectors with the longest norm were selected.

Alpha diversities were computed using phylodiversity metrics (McCoy & Matsen, 2013). The effect of variation in sequencing depth between samples was controlled for by taking average diversity estimates from 100 rarefying draws of the marker gene-matching reads. For each draw, 9,000 reads were considered, which corresponds to the lowest marker gene-matching read count among all samples.

Functional annotation of shotgun metagenomes

We used the metagenomic pipeline of the EBI (Mitchell et al., 2016) to scan reads from the metagenomes with the InterProScan tool for functional protein domain annotation (Mitchell et al., 2014). This analysis was repeated on contigs obtained with Ray assembler (Boisvert, Laviolette, & Corbeil, 2010); as too large a share of the read data was not assembled, we chose to use only the read-based results. Only seven out of the nine Western control datasets were amenable to this analysis as the two samples from (Hasan et al., 2014) have not been publicly released and notably lacked sequencing read quality data. Results are accessible by searching the BioProject accession ERP016024 on the EBI Metagenomics website (https://www.ebi.ac.uk/metagenomics/). We then performed LDA based on the relative abundances of the InterPro terms (normalized by each sample's total annotated read count [Table S1]) to compare HGs to Western controls (Table S5), using a custom R script (http://github.com/flass/microbiomes/tree/master/scripts/interpro/lda_functional.r).

To assess the enrichment of particular biological systems or processes in the different subsistence strategy groups, biological processes that were represented by best-ranking functional terms in the LDA, including pantothenate (vitamin B5) biosynthesis, Coenzyme A related metabolism and urease activity (listed Table S6), had the LDA scores of all their dependent terms compared to those of a control high-ranking process (ribosome). Presence of pantothenate biosynthesis pathway in *Heamophilus* spp. reference genomes was investigated by browsing the Interpro database (www.ebi.ac.uk/interpro, last accessed 11 October 2016).

Results and Discussion

Study design

To circumvent the problem of lack of replication in previous studies on the oral microbiome of hunter-gatherers (HG), we set up a design analyzing three pairs of HG populations and their traditional farmer (TF) neighbors. The three HG populations are the Batak, Aeta and Agta, all members of the 'Negrito' group who are believed to be predominantly descended from the first humans to have settled in the Philippines (Lipson et al., 2014). They live in close proximity with the TF populations, Tagbanua, Zambal and Casigurani, respectively, who are all descendants of a later wave of settlement (Cariño, 2012). The geographic distances between the locations occupied by the pairs of populations range from 1 to 10 km, (Fig. 1; Table S1).

Food exchange between HGs and TFs is common, with up to 50% of the HGs' meals nowadays including rice (Page et al., 2016). Despite this, the two populations maintain distinct diets. HGs are foragers, i.e. still largely relying on fishing, hunting, and gathering (honey, leaves and wild fruits, seeds and tubers; detailed records for Agta Table S7), whereas TFs rely on a traditional farming subsistence strategy, which in the Philippines is mainly based on cultivated rice and vegetables and excludes forest products (Bamberg-Migliano, personal observations). Pairs of HG and TF populations live in close enough proximity to likely be exposed to similar environmental sources of microbes, but their lifestyles differ substantially: the HGs usually live in leantos (without walls), while TFs live in houses with walls; HGs do not brush their teeth, while TFs go to school and receive education relative to oral prophylaxis, and usually have access to toothpaste and a brush (Bamberg-Migliano, personal observations). This setting offers an experimental design of three independent replicates with a comparable level of genetic and ecological differentiation between the populations within pairs, so that any systematic difference in the microbiome species composition between the two groups of populations should be primarily driven by the difference in subsistence strategies and lifestyle.

For each of those populations, we sampled saliva from four individuals in good oral health from which we deep-sequenced the whole extracted DNA, yielding between 167 to 662 million reads per sample, of which 77.0 to 94.7% could be assigned to the human genome (resulting in 2x-20x depth) (Table S1).

Genetic differentiation and admixture between human populations

We explored the genetic structure of the human populations using a robust probabilistic framework suitable for low and variable sequencing depth (Fumagalli, 2013; Skotte, Korneliussen, & Albrechtsen, 2013). Principal component analysis (PCA) and admixture analysis cluster together all individuals from the three TF populations, in accordance with their recent common ancestry (Fig. S2 and S3). However, individuals from the foraging Batak population cluster together with the TFs, while the two other foraging populations form clusters of their own. Only when considering the fourth principal component (PC) of the PCA, or an admixture model with at least four clusters, do the Batak form a cluster of their own that also includes one Tagbanua individual (Fig. S2 and S3). This inference is in line with previous findings based on SNP chips and larger samples (Migliano et al., 2013) and

might be explained by the drastic reduction in population size of the Batak down to 300 individuals in recent times (Scholes et al., 2011). Following this reduction in population size, they developed closer contact with their Tagbanua neighbors, including food trading and occasional marriages (Cariño, 2012), which might have mediated sufficient genetic admixture for them to become more closely related to the farmers.

To ensure that uneven sequencing depth across samples did not affect our estimates of genetic relatedness, we additionally performed a PCA on a sample of the data chosen to equalize the individual population depth. This analysis on the resampled data led to similar patterns of population structure (data not shown) and the principal components did not differ statistically from those obtained from the entire dataset (Procrustes analysis, permutation test, $p \leq 0.0001$) (see Supplementary Methods).

Relative abundances of core oral microbial taxa

The remainder of the reads not mapping to the human genome were used to characterize the composition of oral microbiomes. As an external reference, we included nine methodologically comparable WGS metagenomes of saliva-derived microbiomes from the Human Microbiome Project (HMP) (The Human Microbiome Project Consortium, 2012) and another initiative (Hasan et al., 2014) generated from North-American individuals, hereafter referred to as Westerners. We also attempted to incorporate salivary microbiomes extracted from exome sequencing of South-African HGs (Kidd et al., 2014), but the read depth of those samples was too low to justify their inclusion in the analyses. While we acknowledge that differences in sample collection procedures and sequencing batches may bias the comparison of metagenomes from different datasets, previous studies using datasets of different origins found consistently more similar community compositions between HG groups than between HG and Westerner populations (Clemente et al., 2015; Schnorr et al., 2014), suggesting that batch effects are less important than effects of lifestyle.

We first characterized the microbial composition of all samples using Phylosift (Darling et al., 2014), a pipeline robustly estimating the relative abundances of all lineages of the Tree of Life based on reads matching a dataset of 33 universally conserved marker genes and using a phylogenetic placement framework (Matsen et al., 2010), which accounts evenly for well-characterized clades and deep lineages with few known representatives.

Comparing the Phylosift profiles of our 24 samples of paired HG and TF populations with edge PCA, we found that the PCs of microbiome composition variation in this dataset were driven by differential abundances in widespread oral taxa, including *Veillonella, Streptococcus, Haemophilus, Neisseria, Prevotella*, and various lineages of Actinobacteria. The largest fraction of inter-individual variance in the relative abundance of these taxa (PC1 and PC2 accounting for 52% and 21% of variance, respectively) does not segregate individuals by population or subsistence strategy (Fig. S4 A, B). This suggests that individual factors dominate the major source of variation in oral microbiome composition. This individual noise could reflect the high variation in the oral community within a single host over time, due to regular clearance of microbes by salivary proteins and swallowing of

saliva (Carpenter, 2013). Alternatively, individual differences in nutrition or social relationships involving close physical contact (Kort et al., 2014) may participate in shaping individual microbiomes. However, the following principal components (PC3 and PC4, accounting for 12% and 8% of variance, respectively) result in the separation of the populations by geographic location and subsistence strategy (Fig. S4 C, D). This microbiome composition gradient becomes even more evident when the group of individuals with a Western diet is included in the analysis, as TF populations appear as intermediates between HGs and Western Controls (WCs) (Fig. 2 A, B). The addition of Phylosift estimates of microbiome composition from WC samples to the analysis results in very similar edge PCA plans, regarding the distribution of samples in relation to the vectors of differential abundance (Fig. 2 C; Fig. S4 E, F; Fig. S5 C), and regarding the amount of variance they represent (PC 1-4 respectively account for 47, 18, 14 and 8% of the variance for the 33-sample dataset). Thus, we chose to present the following results in the context of the 33-sample meta-analysis that includes the WC samples; all corresponding graphics and numerical results for the 24-sample analysis – all qualitatively equivalent to those presented below – are available online on the Figshare website at: https://figshare.com/s/e3ba13dbc99a4c87ef25.

Taxonomic gradients reflect geography, host genetics, and subsistence strategy

A first gradient opposing enrichment in *Prevotella* and *Streptococcus* against *Veillonella* and Actinobacteria separates the microbiomes of the three HG populations along the third PC axis (Fig. 2 A, C, E). This could indicate an effect of the local environment, or be linked to genetic differences in the hosts. To distinguish between these hypotheses, we used the reconstructed host genomes associated to the microbiomes to test for correlation of host genetic background and microbiome composition. We computed inter-individual distances in three ways: based on host genotypes (see Sup. Methods); based on their geographic location at time of sampling; and based on the multivariate space depicting their microbiome composition variation. No correlation was observed between Euclidean distances in the full microbiome space and either the host genetic distances or the geographic distances (Mantel test, p-values of 0.53 and 0.62, respectively). However, when considering projections of this multivariate microbiome space on each of its PC, we recovered a trend for an association between partial microbiome distances from the PC3 projection and host genetic distances (Mantel test, $p = 0.078$), as well as a strongly significant correlation with geographic distances (Mantel test, $p = 0.001$). No other PC-projected distances correlated significantly with this factor (Table S2). Host genetics and geography are largely collinear (Mantel test, $p = 0.02$); and after controlling for geography, the correlation between genetic distances and the microbiome-derived PC3 does not remain statistically significant (partial Mantel test, $p = 0.140$). However, after controlling for host genetics, the correlation between geography and microbiome-derived PC3 only slightly decreases (partial Mantel test, $p = 0.002$). Taken together, this suggests that host genetic variation cannot explain microbiome variation on its own, but that the association between geography and microbiome make-up is more robust and likely causal i.e. the local environment, but not host genetics, is likely shaping the composition of the oral microbiome.

This is illustrated by the clustering pattern of samples by geographic origin on PC3 (Fig. 2F). Two out of the three pairs of populations of HGs and farmers share the same mean coordinate on PC3 (Fig. 2C)

(Batak vs. Tagbanua and Aeta vs. Zambal; t-tests p-values of 0.91 and 0.52, respectively). The last pair, however, Agta and Casigurani, shows a marked differentiation on this axis (t-test, $p < 0.005$). This could be explained by the fact that while Agta and Casigurani live in close geographic proximity, their respective villages are separated by an inlet of the sea (Fig. 1), which may contribute to differences in the environments experienced by the two populations. Conversely, the very similar pattern of enrichment in streptococci and *Prevotella* observed for the Batak and Tagbanua could reflect that they often live in the actual same village (Fig. 1), and engage in far more frequent social and genetic exchanges than the other two pairs of populations (Cariño, 2012).

Another composition gradient following the fourth PC axis segregates the samples according to their subsistence strategy (Fig. 2D), with forager populations enriched in *Neisseria* spp. of the *N. lactamica* / *N. meningitidis* / *N. cinerea* group and farmers enriched in *Haemophilus* spp. of the *H. influenzae* / *H. haemolyticus* / *H. aegyptus* group (Fig. 2B, C). This gradient along PC4 appears to be the best way to segregate the subsistence strategies in our sample, as it constitutes the main contributing vector in a linear discriminant analysis of the principal components (DAPC) (Jombart et al., 2010), which results in a very similar projection (Fig. S6), with significant separation between subsistence strategies (Pilai test, $p < 0.030$).

The apparent enrichment of opportunistic pathogens including *N. meningitidis* and *H. influenzae* could be interpreted as being indicative of poor health. However, these species are ubiquitous in the healthy human oral cavity (Costalonga & Herzberg, 2014) and those detected here were likely commensal strains. To further examine this hypothesis, we searched for genes specifically encoding *N. meningitidis* and *H. influenzae* capsular polysaccharides, which are required for virulence. We found limited evidence of their presence in any of the metagenomic assemblies (Supplementary Methods; Table S3, Table S4). This suggests only commensal *Neisseria* spp. and *Haemophilus* spp. that lack established virulence factors colonized the oral cavities of the studied individuals.

Species considered pathogenic discriminate between subsistence strategies

Because increases in the abundance of a few key species can lead to disease (Chen et al., 2015), we also examined fine variation in abundances for all taxa, including rare ones. To do so, we used an alternative method of classification of metagenomic sequences, Kraken, which relies on the finding of exact matches between the metagenomic reads and a large database of complete genomes (Wood & Salzberg, 2014). This method not only provides highly accurate classification of reads, but also makes use of the total information from the WGS dataset and thus provides the best possible estimate of the relative abundances of taxa. We used a linear discriminant analysis (LDA) and LDA effect size (LefSe) (Segata et al., 2011) to find the species with most markedly contrasting abundances across the three lifestyles. The top discriminating taxa included a number of species previously associated with periodontal disease (Chen et al., 2015; Torrungruang et al. 2015): *Prevotella intermedia, Porphyromonas gingivalis, Treponema denticola, Tannerella forsythia, Aggregatibacter actinomycetemcomitans* and *Eubacterium nodatum* were associated with foraging and farming subsistence strategies (Fig. 3). Intriguingly, despite carrying taxa associated with periodontal disease

411 at higher rates than the traditional farmers, the HGs in the Philippines are seemingly in far better oral
412 health (Bamberg-Migliano, pers. obs.).

413 This apparent lack of a negative effect of taxa previously associated with periodontal disease in
414 developed countries on the HGs' oral health suggests these might behave as commensals in HGs and
415 participate in the processing of foods specific to the foragers' diet. This has been previously
416 hypothesized for *Treponema* species in the gut of African and American HGs, which supposedly help
417 degradation of ligneous plant materials (Obregon-Tito et al., 2015; Schnorr et al., 2014). Such
418 commensals may have been present in the ancestral human oral cavity and secondarily lost in
419 populations with increased sanitation and lack of exposure to environmental sources. The only species
420 identified at a markedly higher prevalence in Westerners relative to the other groups is *Cutibacterium*
421 (formerly *Propionibacterium*) *acnes*, an organism mostly associated to skin follicles, but is also found
422 in the digestive tract. At the genus level *Bacteroides, Cutibacterium* and *Campylobacter* also show
423 enrichment in Westerners (significant under ANCOM tests), with the latter genus notably represented
424 by *C. concisus*, a species which abundance is up to 5% of a sample (Fig. 3). *C. concisus* has been
425 hypothesized to be associated with Crohn's disease (Kaakoush et al., 2014), an inflammatory bowel
426 disease with landmark high incidence in the developed world.

427 **Global shifts in species composition**

428 The gradient pattern of *Neisseria* spp. abundances (gradually higher in HGs than in TFs and WCs),
429 seen in the Phylosift-based edge PCA (Fig. 2E, G), is confirmed by Kraken analysis in several species
430 (*N. sicca*, *N. flavescens*, *N. gonorrhoeae,* FDR-corrected Wilcoxon rank-sum test p-value < 0.05) (Fig.
431 3). In contrast, the opposite gradient of *Heamophilus* spp. is not recovered by the Kraken analysis,
432 possibly due to the high variance of estimated abundances for WC samples, ranging between 0-18% of
433 the microbiome composition; the higher prevalence of the *Heamophilus* genus in TFs than in HGs is
434 however confirmed by the Kraken analysis (supplemental data online at:
435 https://figshare.com/s/72d9a99703c222f3ecfb), indicating that the depletion of this taxon in HGs is a
436 robust feature.

437 The enrichment in *Neisseria* spp. in the oral microbiota of HGs versus Westerners was also observed
438 in a comparison between Westerners and South African HGs (Kidd et al., 2014), but Neisseriaceae did
439 not discriminate central African HGs from TFs (Nasidze et al., 2011), and were found depleted in
440 Amerindian HGs relative to Westerners (Clemente et al., 2015). Moreover, all three studies found an
441 enrichment of *Haemophilus* spp. in HGs' saliva. This suggests that the balance between these
442 proteobacterial lineages is an important feature discriminating subsistence strategies, but their relative
443 abundance may still be impacted by additional variables specific to each population. Similarly, an
444 enrichment in Prevotellaceae in HGs, as opposed to an enrichment in *Veillonella* in Westerners, was
445 previously reported (Clemente et al., 2015; Kidd et al., 2014); a similar contrasting microbial
446 enrichment is also observed in our marker gene-based (Phylosift) analysis, but is largely independent
447 of the foragers vs. Westerners divide, and rather characterizes the genetic diversity or geographical
448 location of populations on PC3 (Fig. 2C, E, G). This highlights the importance of controlling for such
449 confounding variables when identifying subsistence strategy-associated oral microbes. At the finer

level, as revealed by our WGS-based (Kraken) analysis, some species of *Prevotella* are indeed enriched in foragers (*P. intermedia* and *P. shahii*), but another lineage (*P.* sp. HMSC077E09) is enriched in Westerners, explaining the absence of lifestyle-discriminating signal at higher taxonomic ranks.

Increased diversity in the oral microbiomes of Hunter-Gatherers

Using the Phylosift framework, we measured the diversity of microbes present in the salivary samples. This revealed a significantly larger phylogenetic diversity (PD) in HGs than in Westerners (t-test, $p < 0.02$), with the Filipino farmers occupying intermediate values (Fig. 4 A), mirroring the gradient observed in relative abundances of core oral taxa (Fig. 2 D). This difference remains significant (t-test, $p < 0.04$), when using balance-weighted PD (BPWD), a measure of diversity partially weighted by lineage abundance (scaling parameter $\Theta = 0.25$) that has been shown to be robust to variation in sampling depth (McCoy & Matsen, 2013). The trend in diversity is also maintained when rarefying all samples to the lowest depth in the dataset (9,000 marker gene-mapped reads), but at this point it loses statistical significance (t-test, $p = 0.07$). Interestingly, this trend emerges from a systematic increase in mean diversity between population of HGs relative to their paired TF population (Fig. 4 C), notwithstanding variations between geographical groups (Fig. 4 B).

An increased diversity in the oral microbiota is generally interpreted as evidence of poorer oral health (Costalonga & Herzberg, 2014). The mouth ecosystem is regularly cleared and re-colonized, and the opening of new niches in gingival crevices and cavities, as well as presence of carbohydrates, can lead to colonization by opportunist microbes and over-growth of commensal taxa into invasive ones (Costalonga & Herzberg, 2014). However, these observations concern individuals living a modern Western lifestyle. In the context of Philippines' HGs, an alternative hypothesis would be that higher diversity is linked to an extended commensal microbiota, possibly leading to gains of function. We therefore investigated in more details what differentiates the taxonomic and functional structures of the microbiomes of each subsistence strategy group.

Functional analysis reveals potential adaptations to diet

Species classification may prove a limited predictor of microbial community function due to phenotypic diversity of bacterial strains within the same species (Zhu, Sunagawa, Mende, & Bork, 2015). We thus used InterProScan to directly annotate metagenomic reads with biochemical functions. From the total of 701,201,172 submitted reads (559,190,367 from the 24 Philippines samples), 242,348,233 coding sequences (136,537,593) were predicted, out of which 76,272,138 (50,384,528) had a functional signature match to InterPro, covering together 11,307 unique functional terms (detailed results accessible at https://www.ebi.ac.uk/metagenomics/projects/ERP016024). We first applied PCA to explore the structure of the functional variation within our dataset. We observed that neither subsistence strategy groups nor populations are well separated along the first six principal components (together accounting to 80% of total variance), indicating that there is no marked functional differentiation between those groups (Fig. S7). However, only a few functions with significant differences abundant functions could still result in relevant ecological differences.

We thus applied LDA to this functional profile, searching for terms that discriminated HGs from Westerners (Table S5). Amongst the top 95 (top 1%) discriminant annotations, we found several terms relating to a few pathways: ribosome structure (11 in the top 1% out of 135 annotations), urease activity (2/9 in top 1%), pantothenate (vitamin B5) and coenzyme A (CoA) biosynthesis (3/9 in top 1%) and CoA-dependent lipid metabolism (5/60 in top 1%) (Table S6). The directions of the imbalances of ribosomal protein-coding sequences were randomly distributed (6 enriched in foragers, 5 enriched in Westerners in the top 1%; 60 and 75 in total), indicating that, when considered globally, ribosome function is evenly distributed, as expected. In contrast, urease annotations were consistently enriched in Westerners (8/9 of all annotations), and CoA-related annotations were consistently enriched in foragers (all of biosynthesis-related annotations and 40/60 of the lipid metabolism-related annotations, including all in the top 1% discriminant ones).

Microbiomes from Westerners, and TFs to a lesser extent, were found to be enriched in metagenomic reads associated with urease function. Reads annotated for this function were mostly assigned to the *Haemophilus* genus (Sup File S2), consistent with our taxononomic abundance-based analysis and with the established ureolytic function of this lineage (Burne & Marquis, 2000). This enzymatic pathway leads to the alkalinogenic release of ammonia, a reaction known to help buffer dental biofilms against acidification. A drop in pH typically occurs when saccharolytic bacteria rapidly degrade free sugars into acidic compounds, promoting tooth demineralization and favoring the growth of cariogenic bacteria (Liu, Nascimento, & Burne, 2012; Reyes et al., 2014). The reduced abundance of *Haemophilus* in HGs' saliva might therefore be expected to lead to dental plaque acidification, and the development of oral diseases like caries. However, this would also require the presence of acidogenic bacteria and more crucially their sugar substrates, which the hunter-gatherer diet is unlikely to provide, as can be seen for the Agta, for whom extensive diet data have been collected (Table S7). This is more likely to happen to the TFs and Westerners, whose diets are richer in starch and processed sugars (Britten et al., 2012). It has been shown in synthetic oral communities repeatedly exposed to pH drops that aciduric species including *Veillonella* spp. increased in frequency and excluded *Neisseria* spp. (Bradshaw & Marsh, 1998), a pattern reminiscent of our observations (Figure 2).

Conversely, we observed an opposite gradient with highest prevalence in foragers of vitamin B5 biosynthetic pathway-associated metagenomic reads. This indicates that microbes autotrophic for this vitamin are more successful at colonizing the mouths of HGs and to a lesser extent TFs. This same trait has been previously observed as the most marked genomic difference between *Campylobacter* spp. colonizing guts of cattle versus poultry. The frequent absence of genes in the vitamin B5 biosynthesis pathway in chicken-associated strains was suggested to reflect their diet of vitamin B5-rich cereals and grains, as opposed to the grass-based cattle diet (Sheppard et al., 2013). Similarly, the difference we observe may reflect the abundance of this essential nutrient in processed food present in the Western diet, as opposed to its scarcity in food consumed by populations from the Philippines. According to daily records of food consumed in Agta camps and in the general American population (see Supplementary Methods), Americans and Agta eat food with globally comparable concentrations of vitamin B5, but Americans have much larger daily portions (Table S8), and hence Westerners consume greater quantities of vitamin B5. This discrepancy in daily ingested quantities of vitamin B5

may result in a different availability of this vitamin in the saliva of each group, which could have impacted the profile of microbes colonizing their oral cavities.

The relative lack of vitamin B5 in the Philippines foragers' diet could select for microbes that are able to synthesize it *de novo*. Such selective pressure on the oral microbiome may explain some of the taxonomic signatures we found to be associated with subsistence strategies. Notably, *Haemophilus* spp., which we showed to be the main bacterial lineage depleted in the HGs' microbiomes (Fig. 2), have genomes devoid of the relevant vitamin B5 biosynthesis pathway (see Supplementary Methods), suggesting *Haemophilus* spp. are counter-selected in the HGs' saliva. Conversely, the diet of Westerners, which provides them with a greater intake of vitamin B5, may have allowed the colonization of the mouth of certain individuals by bacteria auxotrophic for this nutrient, such as *Haemophilus* spp.. An increased abundance of this particular lineage, with its urease activity able to counter acidic bursts, could in turn have geared the microbiome towards an adaptive response to the Westerner's acidogenic sugar-rich diet.

Conclusion

Despite high inter-individual variability and a strong impact of the geographic location of host populations on oral microbiome composition, we were able to recover consistent differentiation associated to subsistence strategies thanks to the replicated design of the study. Key signatures of subsistence strategies include shifts in species distribution including relative abundance of core species such as *Neisseria* spp. vs. *Haemophilus* spp.. This suggests that the hunter-gatherer and traditional farmer diets in themselves, or closely associated ecological or socio-economic factors, are significant drivers of differentiation in saliva.

Our results paint an interesting picture of the oral microbiome in HGs in terms of health and disease. Oral microbiomes from HGs were significantly more diverse than those from TFs or Westerners, as was found previously in distant hunter-gatherer populations (Clemente et al., 2015; Nasidze et al., 2011). While high diversity of microbiomes in the oral cavity has been associated with disease (Griffen et al., 2012), some of this diversity is likely to be adaptive to their forager diet as possibly illustrated by the presence of species involved in the degradation of ligneous material such as *Treponema* spp. (Obregon-Tito et al., 2015; Schnorr et al., 2014). While the HG microbiomes comprise an excess of species that have been shown to be associated to oral disease, it is unclear to what extent these species cause disease in HGs. Indeed, all subjects enrolled in this study were apparently in good oral health and HGs in the Philippines tend to have systematically less caries than the TFs (Bamberg-Migliano, pers. obs.). It is possible that the species complex associated to gingivitis and periodontitis might be part of the healthy microbiota of the HGs' buccal cavity, with pathogenic strains only selected in populations subsisting on a diet richer in starch and refined sugar.

Acknowledgements

The authors acknowledge support from the European Research Council (ERC) (grant ERC260801 – BIG_IDEA), and the National Institute for Health Research University College London Hospitals Biomedical Research Centre. FL was additionally supported by the UK Medical Research Council (grant 412 MR/N010760/1). MS was supported by an Institut Pasteur-Cenci Bolognetti fellowship. MF was supported by a Human Frontiers (HFSP) fellowship. L.S. was supported by a PhD scholarship from EPSRC (EP/F500351/1) and the Reuben Centre for Paediatric Virology and Metagenomics. MD was supported by Leverhulme Trust (grant RP2011-R-045 to ABM), and by French National Research Agency (ANR) grant Labex IAST. We are grateful for comments on the manuscript by Mark Achtman. Finally, we wish to thank the National Commission on Indigenous Peoples (NCIP) and the anonymous donors of saliva samples that were used in the study.

Author Contributions

MS, MGT and FB designed the study; ABM provided samples; MS did the molecular analyses; MS, FL, MF, LS and FB performed the bioinformatics and computational analyses; MS, FL, ABM and FB wrote the paper; ABM, MD, CW and MGT collected and modeled diet information; all authors read and commented on the manuscript.

Conflict of Interests statement

We declare no conflict of interest.

Data availability

- Microbial metagenomic read datasets (after trimming, quality filtering and removal of Kraken-assigned human reads): ENA (www.ebi.ac.uk/ena), BioSample accessions ERS1202862—ERS1202885.

- Results of the EBI metagenome analysis pipeline: EBI Metagenomics (https://www.ebi.ac.uk/metagenomics), project accession ERP016024.

- Output of Kraken (tables of taxon abundances): Figshare, https://figshare.com/s/72d9a99703c222f3ecfb

- Output of guppy (placement edge difference data matrix, phylodiversity estimates and edgePCA projections): Figshare,

https://figshare.com/s/e3ba13dbc99a4c87ef25

- Output of LDA and PCA on Interproscan functional classification: Figshare

https://figshare.com/s/7025b44db131be8caa2f

- Sup. Files S1 and S2: Figshare,

https://figshare.com/s/b46f7cda485add6c0fd7

(temporary Figshare private links FOR REVIEWER USE ONLY; data will be released publicly upon publication)

- Output of Phylosift (placement files) is too bulky (4.7 GB) to be deposited on a data repository personal account, therefore a request to use the journal's sponsored data publishing on DYRAD has been made.

591 **References**

Adler, C. J., Dobney, K., Weyrich, L. S., Kaidonis, J., Walker, A. W., Haak, W., … Cooper, A. (2013). Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions. *Nature Genetics*, *45*(4), 450–455. doi:10.1038/ng.2536

Bocquet-Appel, J.-P. (2011). When the world's population took off: the springboard of the Neolithic Demographic Transition. *Science (New York, N.Y.)*, *333*(6042), 560–561. doi:10.1126/science.1208880

Boisvert, S., Laviolette, F., & Corbeil, J. (2010). Ray: Simultaneous Assembly of Reads from a Mix of High-Throughput Sequencing Technologies. *Journal of Computational Biology*, *17*(11), 1519–1533. doi:10.1089/cmb.2009.0238

Bradshaw, D. J., & Marsh, P. D. (1998). Analysis of pH-driven disruption of oral microbial communities in vitro. *Caries Research*, *32*(6), 456–462.

Britten, P., Cleveland, L. E., Koegel, K. L., Kuczynski, K. J., & Nickols-Richardson, S. M. (2012). Impact of typical rather than nutrient-dense food choices in the US Department of Agriculture Food Patterns. *Journal of the Academy of Nutrition and Dietetics*, *112*(10), 1560–1569. doi:10.1016/j.jand.2012.06.360

Burne, R. A., & Marquis, R. E. (2000). Alkali production by oral bacteria and protection against dental caries. *FEMS Microbiology Letters*, *193*(1), 1–6. doi:10.1111/j.1574-6968.2000.tb09393.x

Cariño, J. K. (2012). Country Technical Notes on Indigenous People's Issues: Republic of the Philippines. Retrieved from https://www.ifad.org/documents/10180/0c348367-f9e9-42ec-89e9-3ddbea5a14ac

Carpenter, G. H. (2013). The Secretion, Components, and Properties of Saliva. *Annual Review of Food Science and Technology*, *4*(1), 267–276. doi:10.1146/annurev-food-030212-182700

Chen, H., Liu, Y., Zhang, M., Wang, G., Qi, Z., Bridgewater, L., … Pang, X. (2015). A Filifactor alocis-centered co-occurrence group associates with periodontitis across different oral habitats. *Scientific Reports*, *5*. doi:10.1038/srep09053

Clemente, J. C., Pehrsson, E. C., Blaser, M. J., Sandhu, K., Gao, Z., Wang, B., … Dominguez-Bello, M. G. (2015). The microbiome of uncontacted Amerindians. *Science Advances*, *1*(3), e1500183–e1500183. doi:10.1126/sciadv.1500183

Cordain, L., Eaton, S. B., Sebastian, A., Mann, N., Lindeberg, S., Watkins, B. A., … Brand-Miller, J. (2005). Origins and evolution of the Western diet: health implications for the 21st century. *The American Journal of Clinical Nutrition*, *81*(2), 341–354.

Costalonga, M., & Herzberg, M. C. (2014). The oral microbiome and the immunobiology of periodontal disease and caries. *Immunology Letters*, *162*(2, Part A), 22–38. doi:10.1016/j.imlet.2014.08.017

Darling, A. E., Jospin, G., Lowe, E., Matsen, F. A., Bik, H. M., & Eisen, J. A. (2014). PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ*, *2*, e243. doi:10.7717/peerj.243

Dewhirst, F. E., Chen, T., Izard, J., Paster, B. J., Tanner, A. C. R., Yu, W.-H., … Wade, W. G. (2010). The Human Oral Microbiome. *Journal of Bacteriology*, *192*(19), 5002–5017. doi:10.1128/JB.00542-10

Dufour, A.-B., & Dray, S. (2007). The ade4 Package: Implementing the Duality Diagram for Ecologists. *Journal of Statistical Software*, *22*(i04). Retrieved from https://www.jstatsoft.org/article/view/v022i04

Fumagalli, M. (2013). Assessing the Effect of Sequencing Depth and Sample Size in Population Genetics Inferences. *PLoS ONE*, *8*(11), e79667. doi:10.1371/journal.pone.0079667

Griffen, A. L., Beall, C. J., Campbell, J. H., Firestone, N. D., Kumar, P. S., Yang, Z. K., … Leys, E. J. (2012). Distinct and complex bacterial profiles in human periodontitis and health revealed by 16S pyrosequencing. *The ISME Journal*, *6*(6), 1176–1185. doi:10.1038/ismej.2011.191

Hasan, N. A., Young, B. A., Minard-Smith, A. T., Saeed, K., Li, H., Heizer, E. M., … Colwell, R. R. (2014). Microbial Community Profiling of Human Saliva Using Shotgun Metagenomic Sequencing. *PLoS ONE*, *9*(5), e97699. doi:10.1371/journal.pone.0097699

Hunter, P. (2014). Pulling teeth from history: DNA from ancient teeth can help to yield information about our ancestors' health, diet and diseases. *EMBO Reports*, *15*(9), 923–925. doi:10.15252/embr.201439353

International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, *431*(7011), 931–945. doi:10.1038/nature03001

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, *11*, 94. doi:10.1186/1471-2156-11-94

Kaakoush, N. O., Castaño-Rodríguez, N., Day, A. S., Lemberg, D. A., Leach, S. T., & Mitchell, H. M. (2014). Campylobacter concisus and exotoxin 9 levels in paediatric patients with Crohn's disease and their association with the intestinal microbiota. *Journal of Medical Microbiology*, *63*(1), 99–105. doi:10.1099/jmm.0.067231-0

Kidd, J. M., Sharpton, T. J., Bobo, D., Norman, P. J., Martin, A. R., Carpenter, M. L., … Henn, B. M. (2014). Exome capture from saliva produces high quality genomic and metagenomic data. *BMC Genomics*, *15*(1), 262. doi:10.1186/1471-2164-15-262

Kort, R., Caspers, M., Graaf, A. van de, Egmond, W. van, Keijser, B., & Roeselers, G. (2014). Shaping the oral microbiota through intimate kissing. *Microbiome*, *2*(1), 41. doi:10.1186/2049-2618-2-41

Kultima, J. R., Sunagawa, S., Li, J., Chen, W., Chen, H., Mende, D. R., … Bork, P. (2012). MOCAT: A Metagenomics Assembly and Gene Prediction Toolkit. *PLoS ONE*, *7*(10), e47656. doi:10.1371/journal.pone.0047656

Kuo, L.-C., Polson, A. M., & Kang, T. (2008). Associations between periodontal diseases and systemic diseases: a review of the inter-relationships and interactions with diabetes, respiratory diseases, cardiovascular diseases and osteoporosis. *Public Health*, *122*(4), 417–433. doi:10.1016/j.puhe.2007.07.004

Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., … Venter, J. C. (2007). The Diploid Genome Sequence of an Individual Human. *PLoS Biol*, *5*(10), e254. doi:10.1371/journal.pbio.0050254

Li, R., Yu, C., Li, Y., Lam, T.-W., Yiu, S.-M., Kristiansen, K., & Wang, J. (2009). SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, *25*(15), 1966–1967. doi:10.1093/bioinformatics/btp336

Lipson, M., Loh, P.-R., Patterson, N., Moorjani, P., Ko, Y.-C., Stoneking, M., … Reich, D. (2014). Reconstructing Austronesian population history in Island Southeast Asia. *Nature Communications*, *5*, 4689. doi:10.1038/ncomms5689

Liu, Y.-L., Nascimento, M., & Burne, R. A. (2012). Progress toward understanding the contribution of alkali generation in dental biofilms to inhibition of dental caries. *International Journal of Oral Science*, *4*(3), 135–140. doi:10.1038/ijos.2012.54

Lloyd-Price, J., Abu-Ali, G., & Huttenhower, C. (2016). The healthy human microbiome. *Genome Medicine*, *8*, 51. doi:10.1186/s13073-016-0307-y

Mandal, S., Van Treuren, W., White, R. A., Eggesbø, M., Knight, R., & Peddada, S. D. (2015). Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial Ecology in Health and Disease*, *26*. doi:10.3402/mehd.v26.27663

Marsh, P. D. (2003). Are dental diseases examples of ecological catastrophes? *Microbiology (Reading, England)*, *149*(Pt 2), 279–294. doi:10.1099/mic.0.26082-0

Marsh, P. D., Do, T., Beighton, D., & Devine, D. A. (2016). Influence of saliva on the oral microbiota. *Periodontology 2000*, *70*(1), 80–92. doi:10.1111/prd.12098

Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S. A., … Reich, D. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*, *528*(7583), 499–503. doi:10.1038/nature16152

Matsen, F. A., Kodner, R. B., & Armbrust, E. V. (2010). pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, *11*(1), 538. doi:10.1186/1471-2105-11-538

McCoy, C. O., & Matsen, F. A. (2013). Abundance-weighted phylogenetic diversity measures distinguish microbial community states and are robust to sampling depth. *PeerJ*, *1*, e157. doi:10.7717/peerj.157

Migliano, A. B., Romero, I. G., Metspalu, M., Leavesley, M., Pagani, L., Antao, T., … Kivisild, T. (2013). Evolution of the Pygmy Phenotype: Evidence of Positive Selection from Genome-wide Scans in African, Asian, and Melanesian Pygmies. *Human Biology*, *85*(1–3), 251–284. doi:10.3378/027.085.0313

Mira, A., Pushker, R., & Rodríguez-Valera, F. (2006). The Neolithic revolution of bacterial genomes. *Trends in Microbiology*, *14*(5), 200–206. doi:10.1016/j.tim.2006.03.001

Mitchell, A., Bucchini, F., Cochrane, G., Denise, H., Hoopen, P. ten, Fraser, M., … Finn, R. D. (2016). EBI metagenomics in 2016 - an expanding and evolving resource for the analysis and archiving of metagenomic data. *Nucleic Acids Research*, *44*(D1), D595–D603. doi:10.1093/nar/gkv1195

Mitchell, A., Chang, H.-Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., … Finn, R. D. (2014). The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Research*, *43*(D1), D213–D221. doi:10.1093/nar/gku1243

Morton, E. R., Lynch, J., Froment, A., Lafosse, S., Heyer, E., Przeworski, M., … Ségurel, L. (2015). Variation in Rural African Gut Microbiota Is Strongly Correlated with Colonization by Entamoeba and Subsistence. *PLoS Genetics*, *11*(11), e1005658. doi:10.1371/journal.pgen.1005658

Nasidze, I., Li, J., Schroeder, R., Creasey, J. L., Li, M., & Stoneking, M. (2011). High Diversity of the Saliva Microbiome in Batwa Pygmies. *PLoS ONE*, *6*(8), e23352. doi:10.1371/journal.pone.0023352

Obregon-Tito, A. J., Tito, R. Y., Metcalf, J., Sankaranarayanan, K., Clemente, J. C., Ursell, L. K., … Lewis, C. M. (2015). Subsistence strategies in traditional societies distinguish gut microbiomes. *Nature Communications*, *6*, 6505. doi:10.1038/ncomms7505

Page, A. E., Viguier, S., Dyble, M., Smith, D., Chaudhary, N., Salali, G. D., … Migliano, A. B. (2016). Reproductive trade-offs in extant hunter-gatherers suggest adaptive mechanism for the Neolithic expansion. *Proceedings of the National Academy of Sciences*, *113*(17), 4694–4699. doi:10.1073/pnas.1524031113

Petersen, P. E. (2005). Priorities for research for oral health in the 21st century--the approach of the WHO Global Oral Health Programme. *Community Dental Health*, *22*(2), 71–74.

Quercia, S., Candela, M., Giuliani, C., Turroni, S., Luiselli, D., Rampelli, S., … Pirazzini, C. (2014). From lifetime to evolution: timescales of human gut microbiota adaptation. *Frontiers in Microbiology*, *5*, 587. doi:10.3389/fmicb.2014.00587

Reyes, E., Martin, J., Moncada, G., Neira, M., Palma, P., Gordan, V., … Yevenes, I. (2014). Caries-free subjects have high levels of urease and arginine deiminase activity. *Journal of Applied Oral Science: Revista FOB*, *22*(3), 235–240.

Schnorr, S. L., Candela, M., Rampelli, S., Centanni, M., Consolandi, C., Basaglia, G., … Crittenden, A. N. (2014). Gut microbiome of the Hadza hunter-gatherers. *Nature Communications*, *5*. doi:10.1038/ncomms4654

Scholes, C., Siddle, K., Ducourneau, A., Crivellaro, F., Järve, M., Rootsi, S., … Migliano, A. B. (2011). Genetic diversity and evidence for population admixture in Batak Negritos from Palawan. *American Journal of Physical Anthropology*, *146*(1), 62–72. doi:10.1002/ajpa.21544

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., & Huttenhower, C. (2011). Metagenomic biomarker discovery and explanation. *Genome Biology*, *12*, R60. doi:10.1186/gb-2011-12-6-r60

Sheppard, S. K., Didelot, X., Meric, G., Torralbo, A., Jolley, K. A., Kelly, D. J., … Falush, D. (2013). Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in Campylobacter. *Proceedings of the National Academy of Sciences*, *110*(29), 11923–11927. doi:10.1073/pnas.1305559110

Skotte, L., Korneliussen, T. S., & Albrechtsen, A. (2013). Estimating individual admixture proportions from next generation sequencing data. *Genetics*, *195*(3), 693–702. doi:10.1534/genetics.113.154138

The Human Microbiome Project Consortium. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, *486*(7402), 207–214. doi:10.1038/nature11234

Torrungruang, K., Jitpakdeebordin, S., Charatkulangkun, O., & Gleebbua, Y. (2015). Porphyromonas gingivalis, Aggregatibacter actinomycetemcomitans, and Treponema denticola / Prevotella intermedia Co-Infection Are Associated with Severe Periodontitis in a Thai Population. *PLOS ONE*, *10*(8), e0136646. doi:10.1371/journal.pone.0136646

Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., … Wang, J. (2008). The diploid genome sequence of an Asian individual. *Nature*, *456*(7218), 60–65. doi:10.1038/nature07484

Warinner, C., Rodrigues, J. F. M., Vyas, R., Trachsel, C., Shved, N., Grossmann, J., … Cappellini, E. (2014). Pathogens and host immunity in the ancient human oral cavity. *Nature Genetics*, *46*(4), 336–344. doi:10.1038/ng.2906

Whitmore, S. E., & Lamont, R. J. (2014). Oral bacteria and cancer. *PLoS Pathogens*, *10*(3), e1003933. doi:10.1371/journal.ppat.1003933

Wood, D. E., & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, *15*(3), R46. doi:10.1186/gb-2014-15-3-r46

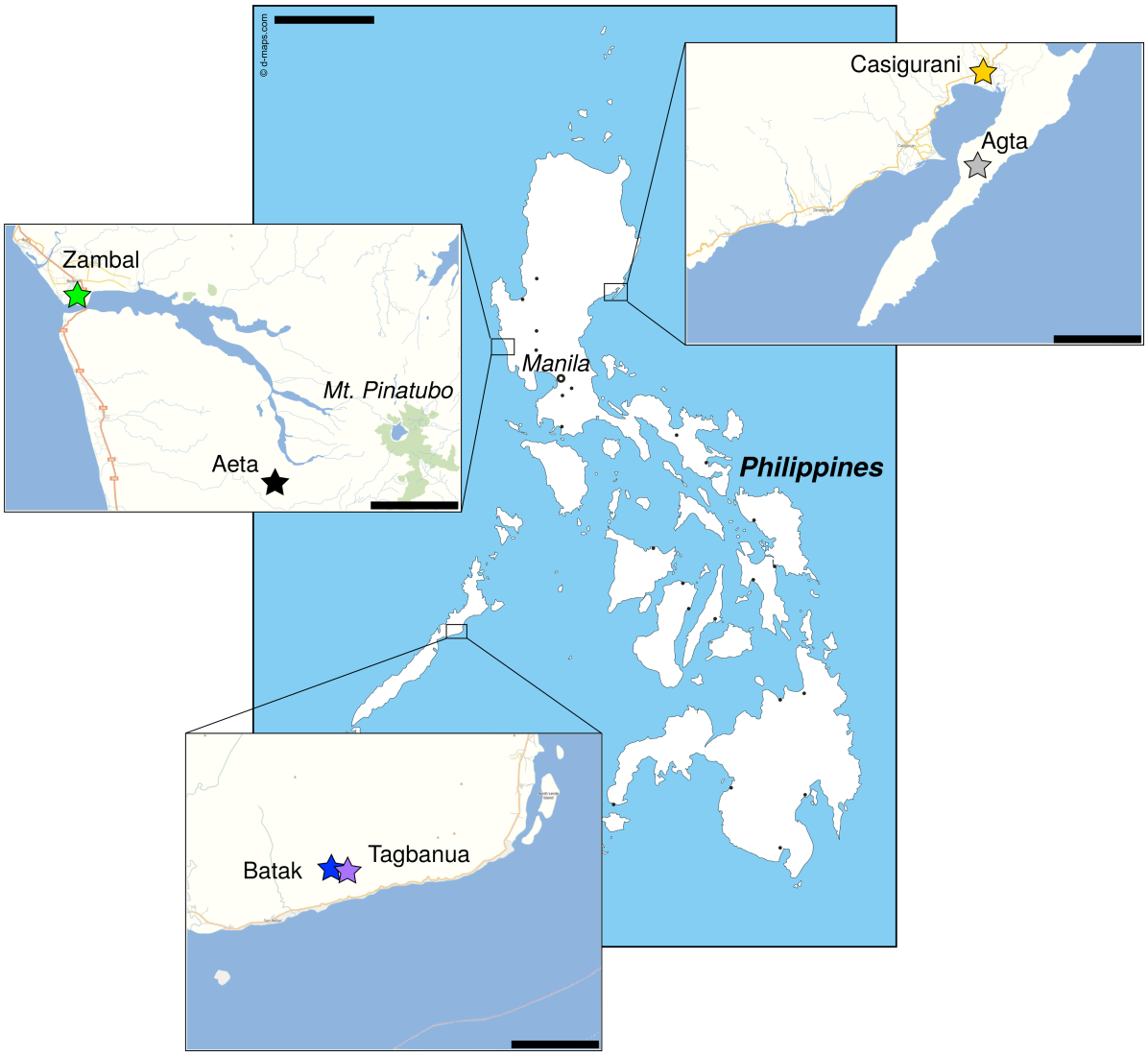
Yang, F., Zeng, X., Ning, K., Liu, K.-L., Lo, C.-C., Wang, W., … Xu, J. (2012). Saliva microbiomes distinguish caries-active from healthy human populations. *The ISME Journal*, *6*(1), 1–10. doi:10.1038/ismej.2011.71

Zaura, E., Nicu, E. A., Krom, B. P., & Keijser, B. J. F. (2014). Acquiring and maintaining a normal oral microbiome: current perspective. *Frontiers in Cellular and Infection Microbiology*, *4*, 85. doi:10.3389/fcimb.2014.00085

Zhu, A., Sunagawa, S., Mende, D. R., & Bork, P. (2015). Inter-individual differences in the gene content of human gut bacterial species. *Genome Biology*, *16*, 82. doi:10.1186/s13059-015-0646-9

592

Figures



593 **Figure 1: Map of the Philippines with location of the study populations**. The three insets highlight
594 the locations of three pairs of populations. Black scale bars represent 200km in the general Philippines
595 map, and 10km in the insets. Maps design obtained from d-maps.com (http://www.d-
596 maps.com/carte.php?num_car=5604&lang=en), LibreMap and OpenStreetMap.

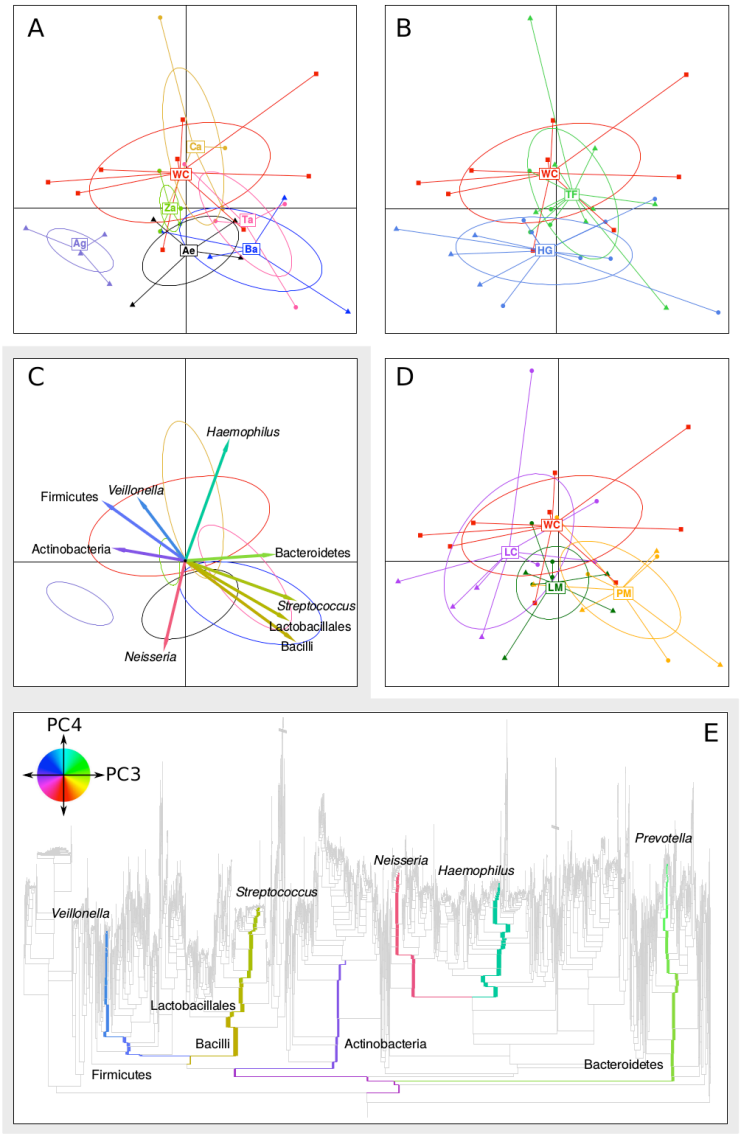


Figure 2: Edge PCA of abundance-weighted microbiome compositions Principal Component Analysis based on Phylosift placement data from the 33-sample dataset. A, B. PC3 and PC4 (x and y axis, 14% and 8% total variance, respectively) projections of variation in lineage abundances across individuals, grouped by population (A) or subsistence strategy (B) C. Same PC3+4 projection highlighting the main contributing variables (lineages of the Tree of Life); ellipses for population groups are represented ghosted in the background. Ellipses represent inertia (variance) of the groups (radius is one time the variance). D. Same PC3+4 projection grouping individual by geographic location. E. Reference Tree of Life on which the major lineages accounting for the variation on PC3+4 are highlighted in colors matching those represented on the plot in (C). Abbreviations: Ae, Aeta; Ag, Agta; Ba, Batak; Ca, Casigurani; Ta, Tagbanua; Za, Zambal; WC, Western Controls; HG, Hunter-Gatherers, TF, Traditional Farmers; LC, Luzon coast; LM, Luzon mountains; PM, Palawan mountains.

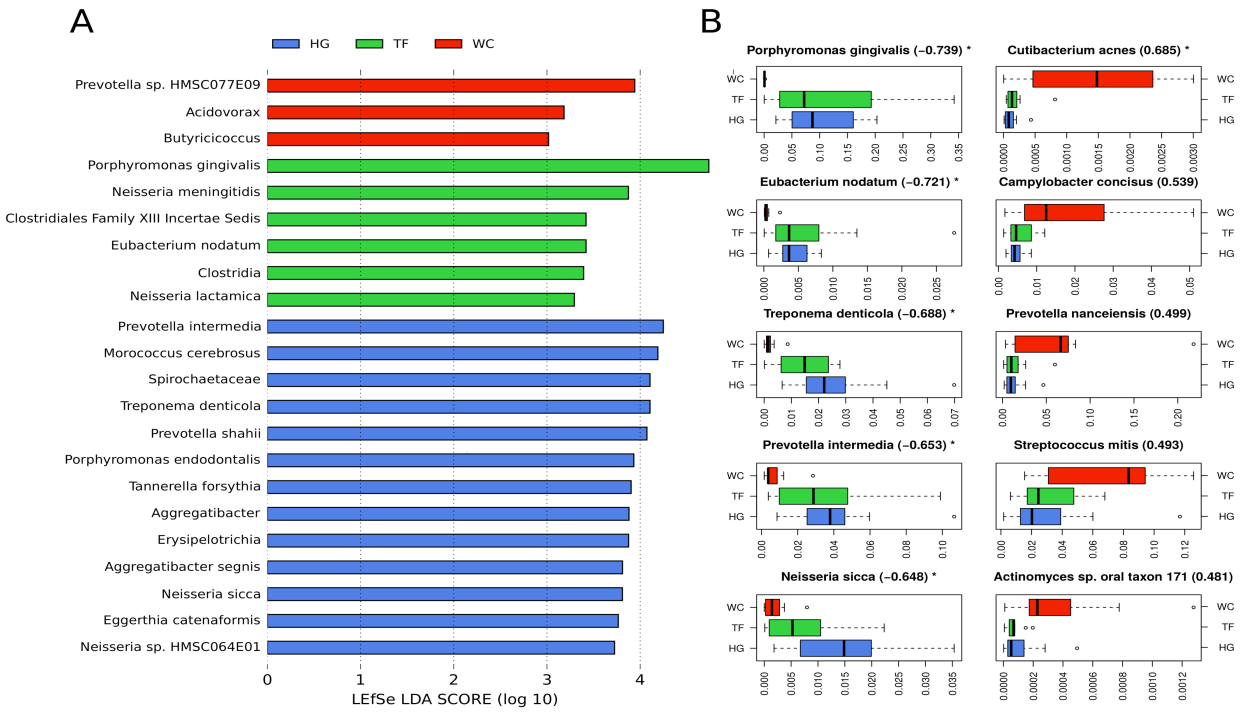


Figure 3: Taxa discriminating between subsistence strategies.
WGS-based estimates of taxonomic abundance (Kraken classification with assignment confidence over 20%) were used to find (A) the best discriminant taxa based on a three-way comparison of the HG, TF and WC groups with the LEfSe algorithm (non-redundant taxa with score over 3 are presented), and (B) the best discriminant species between HG and WC groups based on a simple LDA: left column, species enriched in HGs; right column, species enriched in WCs. Abundances significantly different under a Wilcoxon rank sum test with FDR-corrected p-values < 0.05 are indicated with an asterisk. WC: Western Controls; TF: Traditional Farmers; HG: Hunter-Gatherers. An extended set of the top discriminant taxa is presented online at

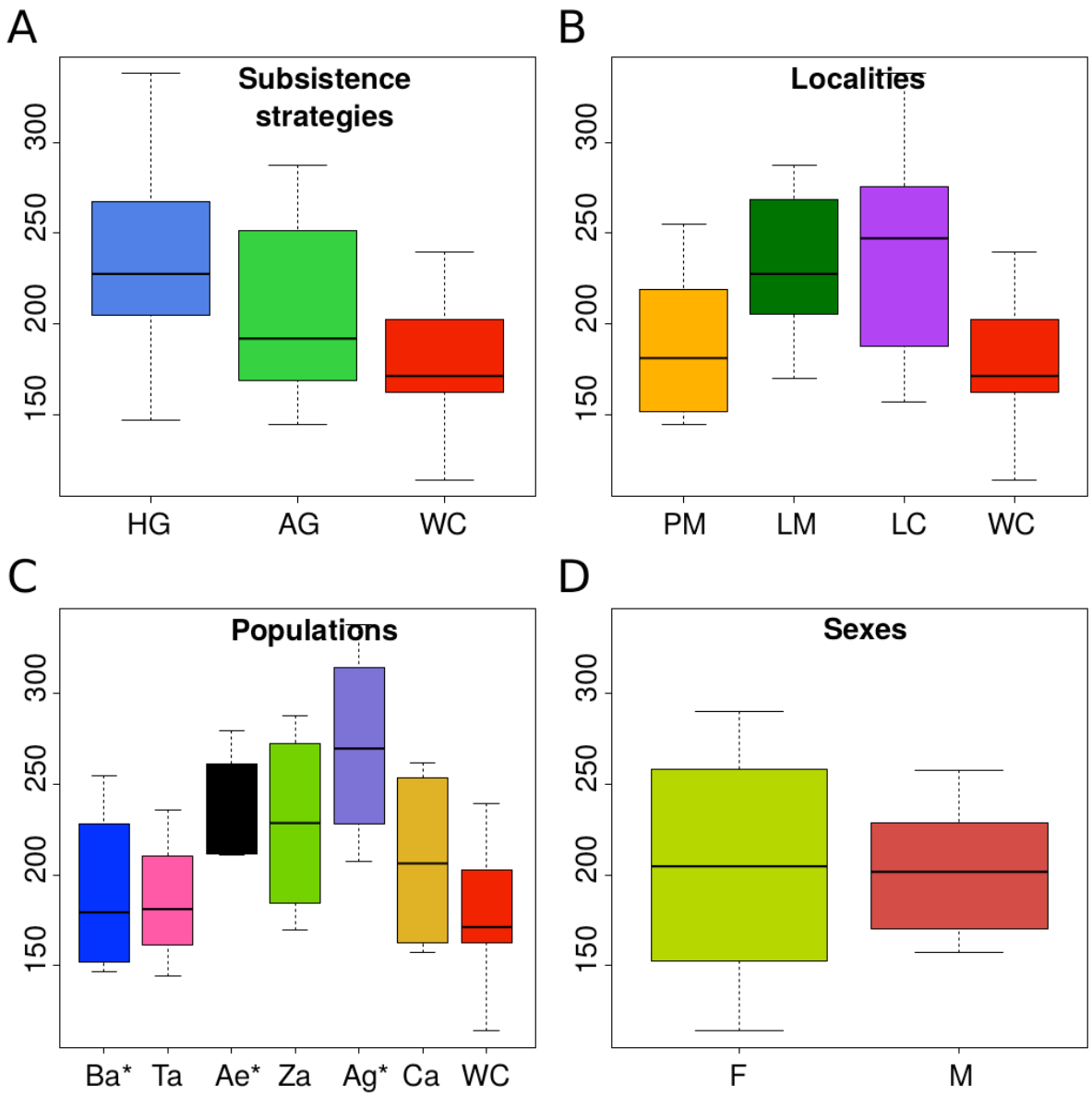


Figure 4: Alpha diversity of metagenomic samples.
Phylogenetic diversity is derived from Phylosift placements and does not consider the relative abundance of lineages. Abbreviations as in Figure 2. HG populations are indicated by an asterisk. Samples are grouped either by subsistence strategy (A), locality (B), sampled population (C) or sex (D).

Supplementary Methods

Supplementary Material

Analysis of human sequencing data

BAM files resulting from the mapping of metagenomic reads on the human genome reference sequence hg19 (Genome Reference Consortium Human Reference [GRCh] 37) using SOAPAligner2 (R. Li et al. 2009) during MOCAT pre-processing (see main Methods) were used for analysis of the human DNA data.

To assess the possible effect of uneven sequencing depth across samples, we downsampled the original BAM files to approximately the average depth of library #2. Specifically, we randomly sampled half of the reads for library #1 and a quarter for library #3 using SAMtools (H. Li et al. 2009).

We performed a Principal Component Analysis (PCA) using ngsTools (Fumagalli et al. 2014), which implements a method based on that of Patterson et al. (2006) but without assigning individual genotypes. This approach has been shown to be more reliable in cases of low or variable sequencing depth (Fumagalli et al. 2013). We filtered out sites where the minimum global depth was below 90 and 40 for the full and sampled dataset, respectively. Likewise, we remove sites with a maximum global depth greater than 310 and 150 for the full and sampled dataset, respectively. These choices allowed for an approximately equal proportion of sites to be filtered out in both datasets, specifically the top 1% and the bottom 5% of the empirical distribution of sequencing depth.
Additionally, we imposed that all samples must have data at each analyzed site and enforced a minimum mapping and base quality score of 20 in Phred score. We used the software ANGSD (Korneliussen et al. 2014) to calculate genotype posterior probabilities using an informative prior under the assumption of Hardy-Weinberg equilibrium across all samples (Kim et al. 2011). We analyzed only putative SNPs with p-value <1e-4 based on a Likelihood Ratio Test (LRT) (Korneliussen et al. 2014). By analyzing chromosome 1, we retrieved 360,301 and 142,878 SNPs for the full and sampled dataset, respectively. These sites were then used to estimate the covariance matrix and perform a PCA. To compare the results obtained from the full and sample data sets, we performed a Procrustes analysis (Wang et al. 2010) on the first four principal components. Significance was assessed by permutations (low *p*-value signifying lower distance statistics than most of the values from the random permutation draws, i.e. supporting closeness of plot shapes). Genetic admixture proportions and pairwise genetic distances were estimated using NGSadmix (Skotte et al. 2013) and ngsDist (Vieira et al. 2015), respectively. These methods again take genotype uncertainty into account, and we used a minimum depth of 30 and a maximum of 300 reads.

This analysis pipeline was implemented in a shell script available at github.com/flass/microbiomes/tree/master/scripts/human/.

Public microbiome data

A total of nine WGS saliva microbiomes from Western individuals have been retrieved from the public databases and used as controls. The available metagenomes in the Sequence Read Archive (SRA) that matched the query "G_DNA_Saliva" AND ILLUMINA" as of July 2014 consisted of seven metagenomes from the HMP project (The Human Microbiome Project Consortium 2012). The corresponding sequence files were downloaded (registered under the BioSample accessions SRS013942, SRS014468, SRS014692, SRS015055, SRS019120, SRS104275 and SRS147126), and two additional metagenomes from another study (Hasan et al. 2014) (BioProject PRJNA231652) were kindly provided after a direct request to the authors of that study. These samples represent to our knowledge the only source of saliva-derived microbiome data where the sequencing coverage was high enough to be used in our study for comparison. We also screened exome data from saliva from Khoisan hunter-gatherers (Kidd et al. 2014), but after filtering and removing human reads, the small number of remaining reads was considered unsuitable for further analysis. Two HMP samples pairs, (SRS014468, SRS015055) and (SRS019120, SRS013942), were generated from the same respective individuals at different time points. The impact of the presence of longitudinal replicates, as well as the general impact of the meta-analysis including third-party samples, was assessed by replicating the edgePCA analysis (see Methods) on restricted datasets: 24 samples including only those from the Philippines generated in the present study, and 31 samples including 7 single-individual HMP samples (removing SRS015055 and SRS013942) and the samples from Hasan et al.

Search for virulence factors

We sought to determine whether the identification of oral pathogens was consistent with the presence of virulence factors for these pathogens. For the forager and farmer samples, microbial reads (i.e. with reads mapping to human filtered out) were assembled into contigs and scaffolds using the Ray assembler (Boisvert et al. 2010) version 2.3.1, using a k-mer length of 31 (-k 31) and default settings for other parameters. These contigs were used to build a nucleotide BLAST database, which was searched for the presence of representative *N. meningitidis* and *H. influenzae* capsular biosynthesis genes. A search against the *N. meningitidis* BIGSdb sequence database capsular scheme (Jolley and Maiden 2010) showed some samples contained sequences with high similarity to some regions of the capsule locus, but no samples contained any hits for region A, the region responsible for capsular biosynthesis (Spinosa et al. 2007). We also searched for similarity to all genes in the partially duplicated capsule locus of *H. influenzae* strain 1007 (GenBank: AF549213.1). Hits for

serotype-specific capsule genes *bcs1* and *bcs2* were only present in 2/24 samples (AE10 and AE12), and this distribution of hits did not appear to be related to sequencing depth.

To assess our power to detect a specific gene within our assembled dataset, we performed two tests. First, we computed the expected frequency g of a gene of a given species in a given sample as follows: we considered the quantities l, the gene size, generically of 1kb; L, the species' genome size (taken as the median reference genome size from NCBI Genome database); a, the species' relative abundance in the sample (as estimated with Kraken, see above); and c, the concatenated length of assembled contigs for the sample (as a proxy of non-redundant genome coverage), to compute:

$$g = a \cdot c \cdot l / L$$

Estimates of gene frequencies for *H. influenzae* and *N. meningitidis* are always greater than one (Table S3).

Second, to empirically test this power to detect genes, we searched for 1kb regions of 16S rRNA genes for each of these species in assembled contigs at 97% sequence identity. We found that we could detect them in the majority of samples (Table S4). Increasing the sequence identity to 99% did not change this conclusion. Both tests thus suggest that we had, in all cases, the power to detect any specific genes from theses pathogen species from our assembled metagenomes, and that virulence-associated capsular genes were likely to be genuinely absent.

Estimation of pantothenic acid intake in American and Agta populations.

Food consumption data were collected between March and July 2014 among six Agta camps located in the municipality of Palanan, Luzon. In each camp, data was collected on consecutive days (mean = 8.5, range 5-10 days). Camps ranged in size from 20 to 114 individuals (mean = 43.8). Two of the study camps were located along the shoreline while the other four were located inland, along rivers. In each camp, data was collected on all foods brought into camp after foraging trips (463 foraging trips in total, mean per observed day = 9.08). When food was returned to camp, the weight of the food was recorded using a Pescola spring scale. If foraged food was traded with local agricultural communities for rice or other products, as frequently occurs among the Agta, the weight of the food received was recorded and was included in dietary calculations. Since food is extensively shared between Agta households (Dyble et al. 2016), it was not always possible to record how much food was consumed by specific individuals. To derive an estimate of nutritional consumption per capita we therefore compiled a total of all foods consumed in each camp during the study period (Table S7). The average dietary profile for Americans was similarly derived from the NHANES 2013-2014 USDA survey (U.S. Department of Agriculture, Agricultural Research

Service 2016) that recoded individual food consumption of 16,166 participants over two days. For both Agta and American diets, the pantothenic acid contents were obtained using USDA National Nutrient Database for Standard Reference, Release 28 (Ahuja et al. 2015). Where no exact match existed in the database, food proxies were chosen in the database based on their phylogenetic closeness to the target foods. Where foods had a non-edible component such as seeds, shells and husks, allowances were made to calculate the edible portion.

The concentration of vitamin B5 in the recorded diets appears similar for Agta (2.14–5.35 μg/g) and Americans (3.26 μg/g) (Table S8), while the ingested quantity is different (1.39–2.67 and 4.38 mg/person/day, respectively). Estimates of per-capita consumption of food – and hence, of vitamin B5 – are, however, likely to be biased by variations in per-capita consumption profiles (e.g. considering variable proportions of infants and adults in populations) and to under-reporting of consumed foods (e.g. non-recording of what was eaten during the foraging activity and not brought back to the foragers camp). The sizes of recorded portions of food (expressed in the energetic value per person per day) indicate these biased estimates are not consistent with real diets. Agta recorded diets range from 440 to 1202 kcal/person/day, which is largely below the daily need of an adult (2,000–2,200 kcal/person/day) or even that of an infant (1,000 kcal/person/day) (Britten et al. 2012), suggesting a significant fraction of the Agta diet was not recorded. Similarly, a survey of the Food and Agriculture Organization of the United Nations (FAO) indicates that the daily dietary energy availability for Americans is 3,750 kcal/person/day (FAOSTAT 2008), suggesting a large under-reporting of consumption in the USDA survey, as been reported previously (Archer et al. 2013). In comparison, the average daily dietary energy availability in the Philippines is 2,580 kcal/person/day, which is likely an overestimate for the forager populations. Hence, the respective reported intake of food and vitamin B5 are both underestimates but are consistently representative of the intake difference between Americans and Agta.

764 # References

Ahuja JKC, Haytowitz D, Pehrsson PR, Roseland J, Exler J, Khan M, Nickle M, Nguyen Q, Patterson K, Showell B, et al. 2015. USDA National Nutrient Database for Standard Reference, Release 28. USDA Available from: http://www.ars.usda

Archer E, Hand GA, Blair SN. 2013. Validity of U.S. Nutritional Surveillance: National Health and Nutrition Examination Survey Caloric Energy Intake Data, 1971–2010. PLOS ONE 8:e76632.

Boisvert S, Laviolette F, Corbeil J. 2010. Ray: Simultaneous Assembly of Reads from a Mix of High-Throughput Sequencing Technologies. J. Comput. Biol. 17:1519–1533.

Britten P, Cleveland LE, Koegel KL, Kuczynski KJ, Nickols-Richardson SM. 2012. Impact of typical rather than nutrient-dense food choices in the US Department of Agriculture Food Patterns. J. Acad. Nutr. Diet. 112:1560–1569.

Dyble M, Thompson J, Smith D, Salali GD, Chaudhary N, Page AE, Vinicuis L, Mace R, Migliano AB. 2016. Networks of Food Sharing Reveal the Functional Significance of Multilevel Sociality in Two Hunter-Gatherer Groups. Curr. Biol. 26:2017–2021.

FAOSTAT. 2008. FAO Food Balance Sheets. Available from: http://www.fao.org/faostat/en/#home

Fumagalli M, Vieira FG, Korneliussen TS, Linderoth T, Huerta-Sánchez E, Albrechtsen A, Nielsen R. 2013. Quantifying Population Genetic Differentiation from Next-Generation Sequencing Data. Genetics 195:979–992.

Fumagalli M, Vieira FG, Linderoth T, Nielsen R. 2014. ngsTools: methods for population genetics analyses from next-generation sequencing data. Bioinformatics 30:1486–1487.

Hasan NA, Young BA, Minard-Smith AT, Saeed K, Li H, Heizer EM, McMillan NJ, Isom R, Abdullah AS, Bornman DM, et al. 2014. Microbial Community Profiling of Human Saliva Using Shotgun Metagenomic Sequencing. PLoS ONE 9:e97699.

Jolley KA, Maiden MC. 2010. BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics 11:595.

Kidd JM, Sharpton TJ, Bobo D, Norman PJ, Martin AR, Carpenter ML, Sikora M, Gignoux CR, Nemat-Gorgani N, Adams A, et al. 2014. Exome capture from saliva produces high quality genomic and metagenomic data. BMC Genomics 15:262.

Kim SY, Lohmueller KE, Albrechtsen A, Li Y, Korneliussen T, Tian G, Grarup N, Jiang T, Andersen G, Witte D, et al. 2011. Estimation of allele frequency and association mapping using next-generation sequencing data. BMC Bioinformatics 12:231.

Korneliussen TS, Albrechtsen A, Nielsen R. 2014. ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics 15:356.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinforma. Oxf. Engl. 25:2078–2079.

Li R, Yu C, Li Y, Lam T-W, Yiu S-M, Kristiansen K, Wang J. 2009. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25:1966–1967.

Patterson N, Price AL, Reich D. 2006. Population Structure and Eigenanalysis. PLoS Genet 2:e190.

Skotte L, Korneliussen TS, Albrechtsen A. 2013. Estimating individual admixture proportions from next generation sequencing data. Genetics 195:693–702.

Spinosa MR, Progida C, Talà A, Cogli L, Alifano P, Bucci C. 2007. The Neisseria meningitidis Capsule Is Important for Intracellular Survival in Human Cells. Infect. Immun. 75:3594–3603.

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

U.S. Department of Agriculture, Agricultural Research Service. 2016. USDA Food and Nutrient Database for Dietary Studies 2013-2014. Available from: http://www.ars.usda.gov/nea/bhnrc/fsrg

Vieira FG, Lassalle F, Korneliussen TS, Fumagalli M. 2015. Improving the estimation of genetic distances from Next-Generation Sequencing data. Biol. J. Linn. Soc. 117:139–149.

Wang C, Szpiech ZA, Degnan JH, Jakobsson M, Pemberton TJ, Hardy JA, Singleton AB, Rosenberg NA. 2010. Comparing spatial maps of human population-genetic variation using Procrustes analysis. Stat. Appl. Genet. Mol. Biol. 9:Article 13.

Supplementary figures

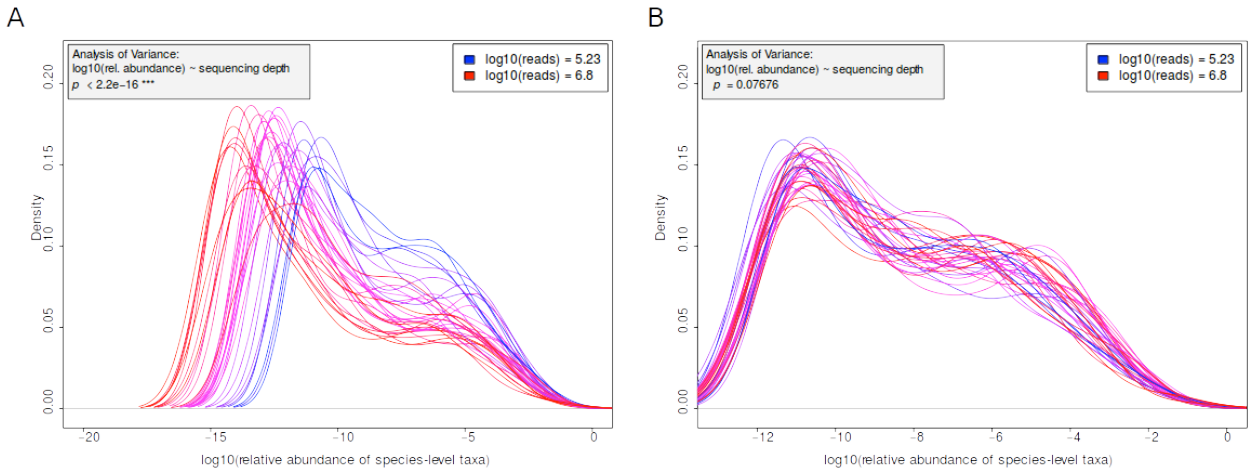


Figure S1: Distribution of relative abundances per sample relative to sequencing depth. Data before (A) and after (B) truncation of the lower-abundance species data. A sample is represented by each curve of species abundance kernel density, which is colored according to its microbiome sequencing depth (the number of reads, excluding those mapping to human), with colors for upper and lower bounds shown in lower insets. Results of an analysis of variance (ANOVA) testing the association of sequencing depth with the distribution of species abundance are shown in the upper insets. The 1e-12 cut-off for the truncated data was chosen as the lower integer exponent value so that the ANOVA Fisher test was not significant.

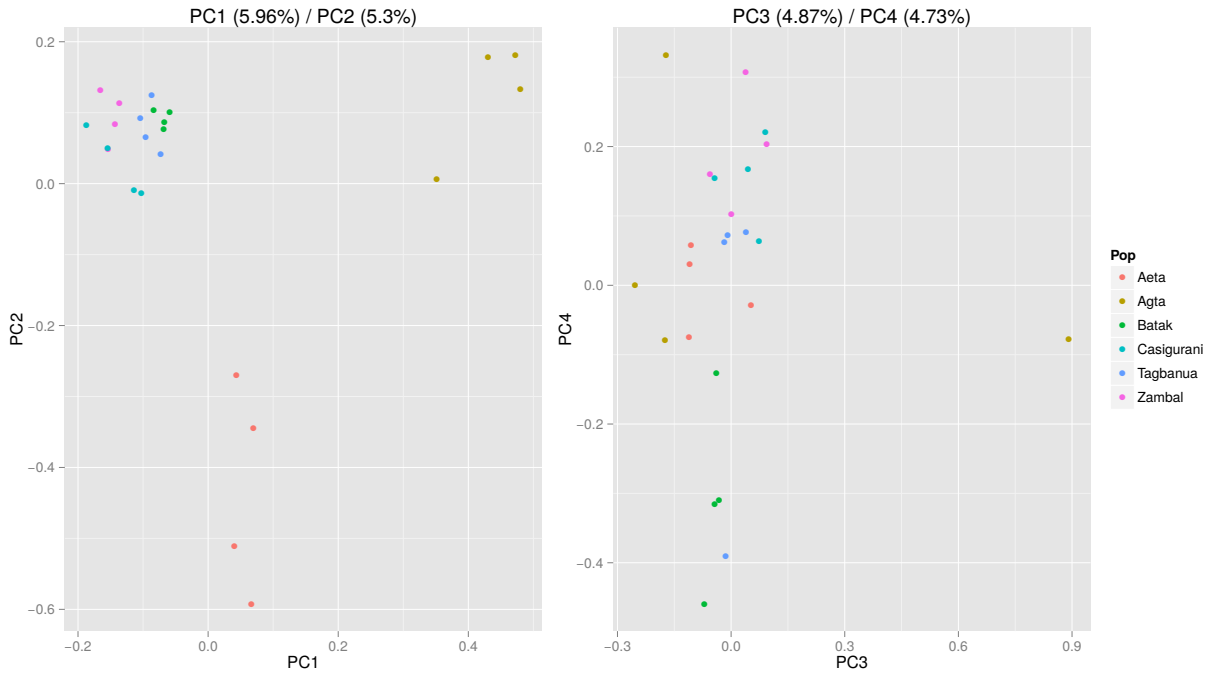


Figure S2. PCA based on covariance matrix of SNPs located on human chromosome 1.

765

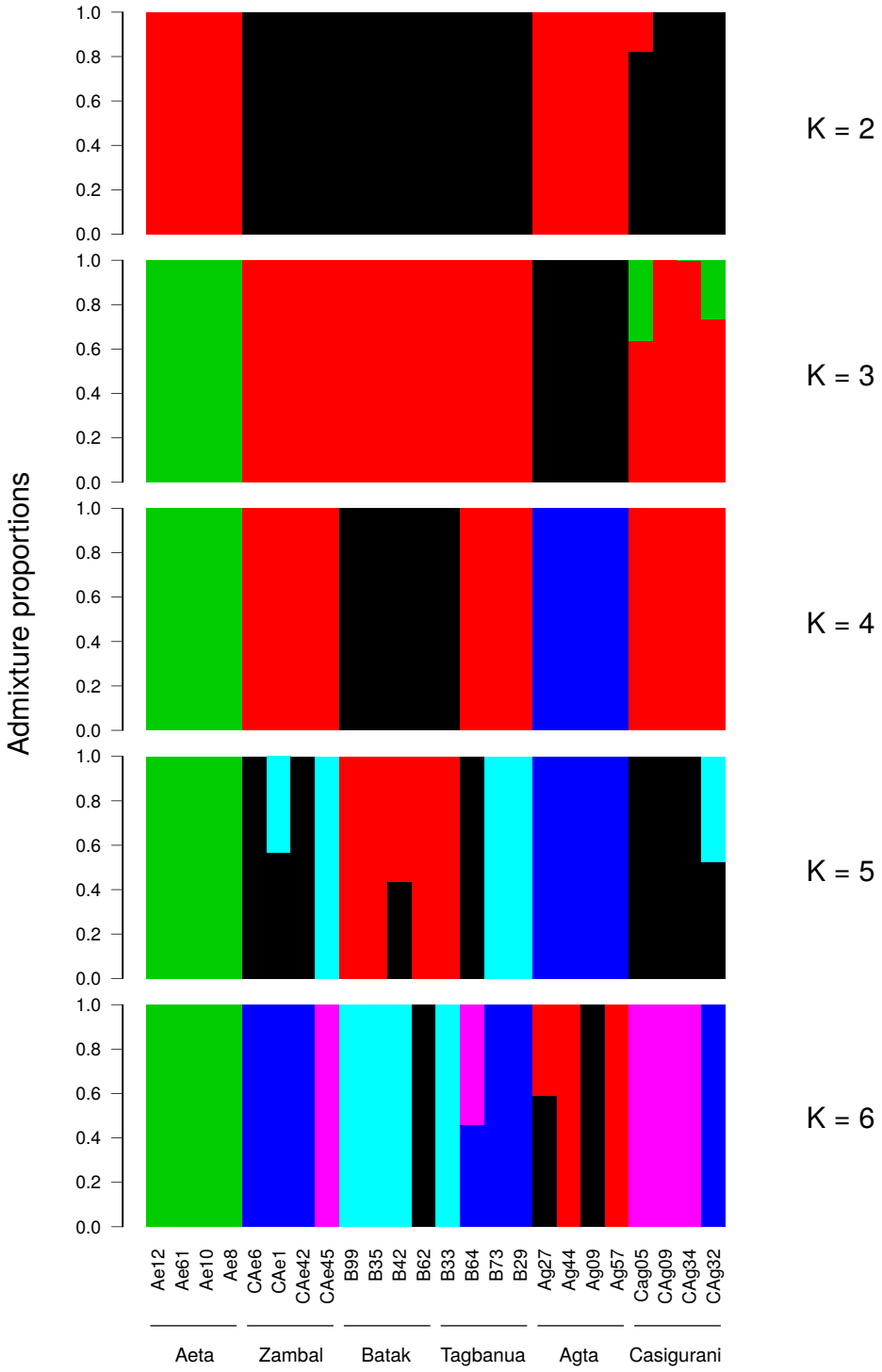


Figure S3. Admixture plots based on SNPs located on human chromosome 1.
Each column represents an individual and the respective proportions of its genotype assigned to either of the K arbitrary clusters (each cluster is represented by an arbitrary color).

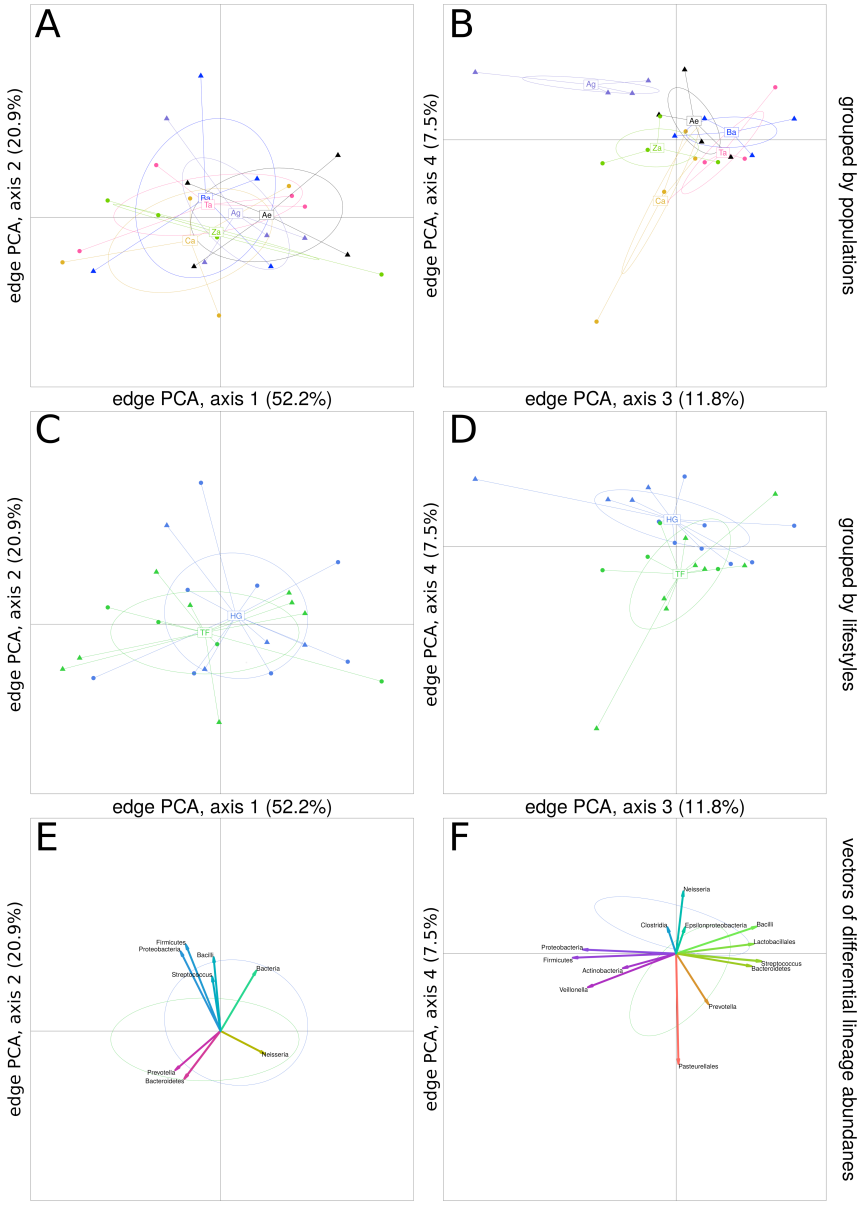


Figure S4: Edge Principal Component Analysis of abundance-weighted microbiome compositions based on Phylosift placement data from the 24-sample WGS dataset.
A, C. PC1 and PC2 (x and y axis, 52% and 21% total variance, respectively) projections of variation in lineage abundances across individuals, grouped by population (A) or subsistence strategy (B). E. Same projection highlighting the main contributing variables (lineages of the Tree of Life); ellipses for subsistence strategy groups are represented ghosted in the background. Ellipses represent inertia (variance) of the groups (radius is one time the variance). B, D, E. Idem, for PC3 and PC4. Abbreviations: Ae, Aeta; Ag, Agta; Ba, Batak; Ca, Casigurani; Ta, Tagbanua; Za, Zambal; WC, Western Controls; HG, Hunter-Gatherers, TF, Traditional Farmers; LC, Luzon coast; LM, Luzon mountains; PM, Palawan mountains.

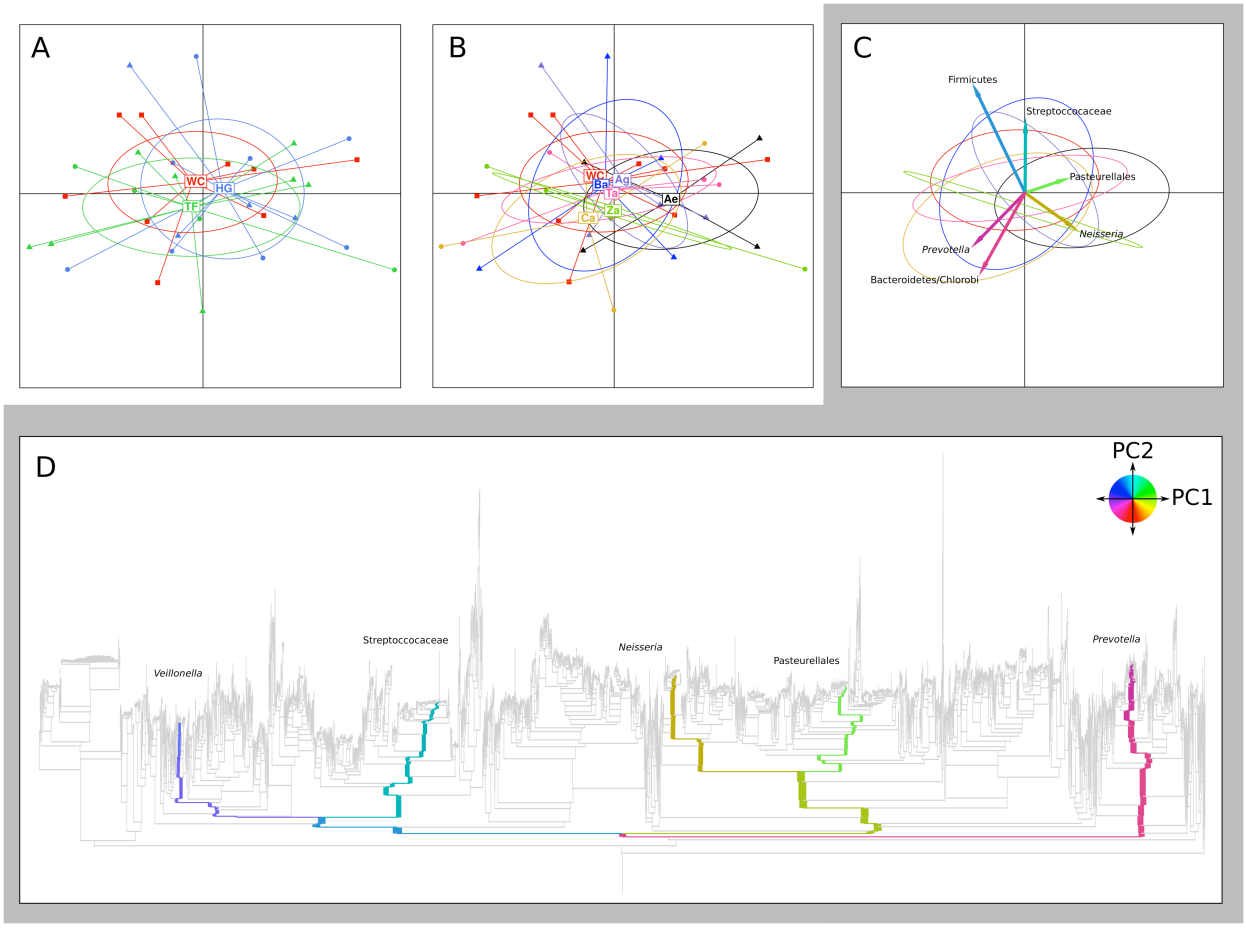


Fig S5: Edge Principal Component Analysis of abundance-weighted microbiome compositions based on Phylosift placement data from the 33-sample WGS dataset. A, B. PC1 and PC2 (x and y axis, 47% and 18% total variance, respectively) projections of variation in lineage abundances across individuals, grouped by population (A) or subsistence strategy (B). C. Same projection highlighting the main contributing variables (lineages of the Tree of Life); ellipses for population groups are represented ghosted in the background. Ellipses represent inertia (variance) of the groups (radius is one time the variance). D. Reference Tree of Life on which the major lineages accounting for the variation on PC1+2 are highlighted in colors matching those represented on the plot in (E). Abbreviations: Ae, Aeta; Ag, Agta; Ba, Batak; Ca, Casigurani; Ta, Tagbanua; Za, Zambal; WC, Western Controls; HG, Hunter-Gatherers, TF, Traditional Farmers; LC, Luzon coast; LM, Luzon mountains; PM, Palawan mountains.

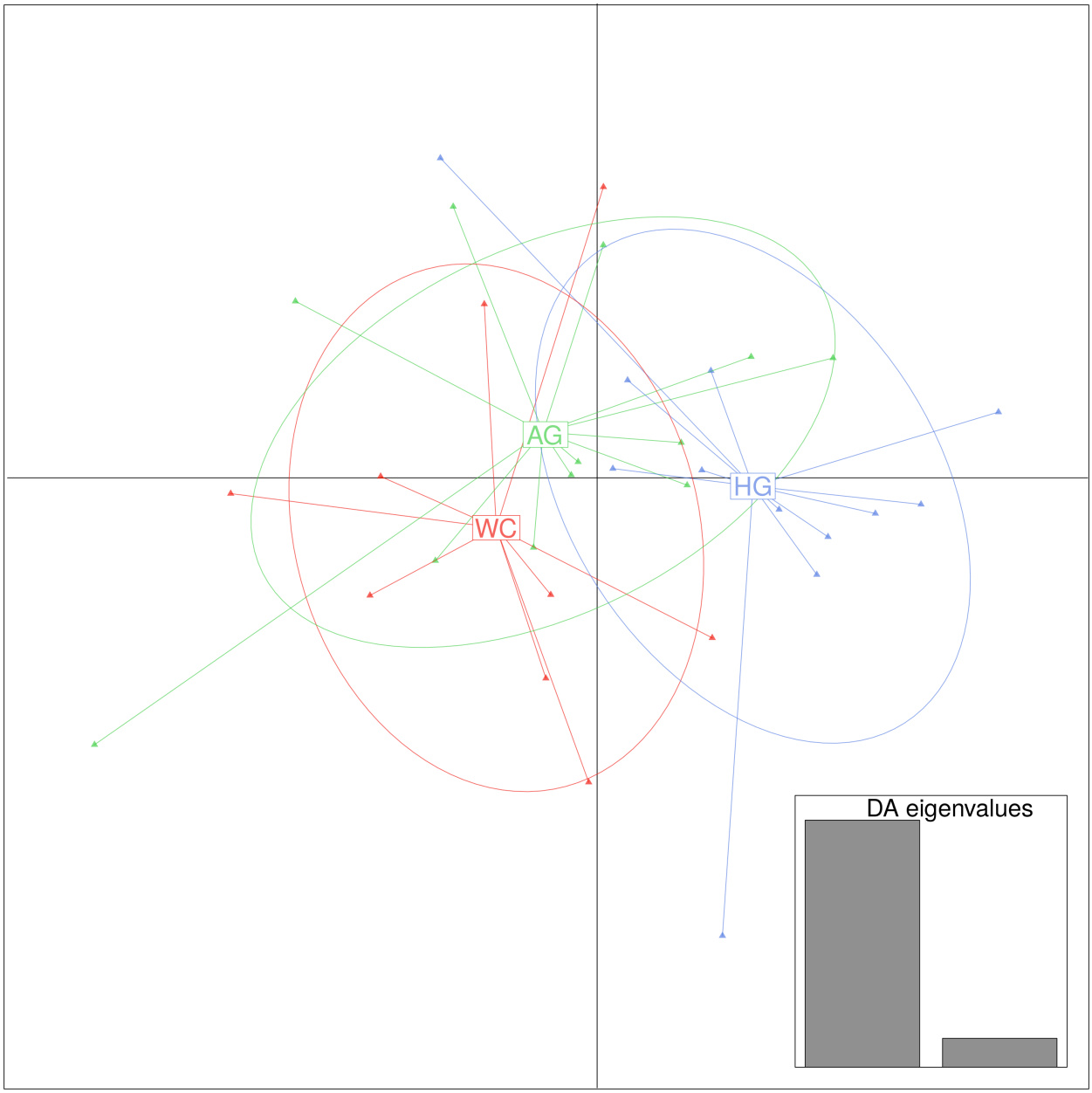
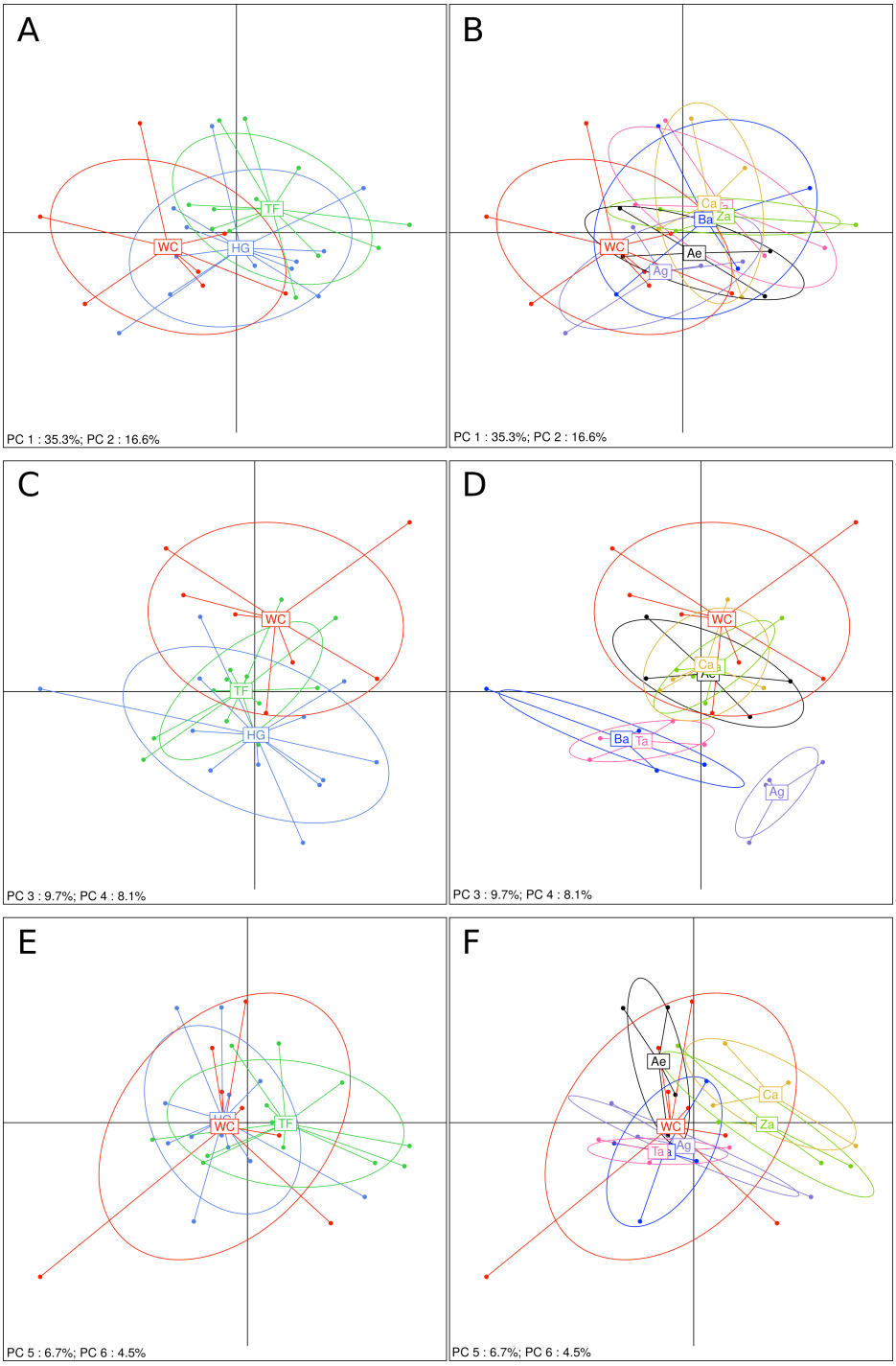


Figure S6: Discriminant Analysis of the Principal Components based on Phylosift placement data generated from 33 WGS samples.

Based on the 4 first PCs of the edge PCA, two discriminant functions were used to separate HG, AG and WC groups of samples. Inset barplot show the relative contribution of discriminant functions on x and y axes to inter-group variance.



784

785 **Figure S7: PCA of relative abundances of InterPro terms**

786 (A,B) PC1 (x axis) and PC2 (y axis); (C,D) PC3 (x axis) and PC4 (y axis); (E,F) PC5 (x axis) and PC6
787 (y axis). (A,C,E) samples grouped by subsistence strategy; (B,D,F) samples grouped by population.