# From sporulation to village differentiation: the shaping of the social microbiome over rural-to-urban lifestyle transition in Indonesia

Clarissa A. Febinia[1,2,#], Hirzi Luqman[1,#,*], Pradiptajati Kusuma[2], Lidwina Priliani[2], Joseph Lewis[1], Desak Made Wihandani[3], Gde Ngurah Indraguna Pinatih[3], Herawati Sudoyo[2], Alexandre Almeida[4], Safarina G. Malik[2,*], Guy S. Jacobs[1,*]

[1] Department of Archaeology, University of Cambridge, Cambridge CB2 3DZ, United Kingdom
[2] Genome Diversity and Disease Division, Mochtar Riady Institute for Nanotechnology, Banten 15811, Indonesia
[3] Faculty of Medicine, Udayana University, Denpasar, Bali 80232, Indonesia
[4] Department of Veterinary Medicine, University of Cambridge, Cambridge CB3 0ES, United Kingdom
[#] These authors contributed equally
* Corresponding authors: gsj22@cam.ac.uk, hl636@cam.ac.uk, safarina.malik@mrinstitute.org

## Summary
Despite established roles in human health and profound global diversity, existing gut microbiome datasets are biased toward Western urban cohorts, with especial under-representation of Southeast Asia. Here, we present a novel gut microbiome dataset from 116 Indonesian individuals representing a diverse cline from transitional hunter-gatherers to rural agricultural to urban lifestyles. We identify 1,304 species and 3,258 subspecies by assembling 11,070 metagenome-assembled genomes, revealing substantial species (15%) and subspecies-level (50%) novelty. Novel taxa are rare, often village-specific, and depleted for sporulation genes, revealing a direct link between bacterial physiology, transmission, prevalence and discovery. We reveal reduced microbial diversity associated with rural-to-urban transition, and also multi-layered patterns across taxonomic levels, including opposing trends within genera. Furthermore, community membership, over diet, is often a stronger predictor of microbiome composition. Our work highlights the interplay of host behaviour, population structure and bacterial physiology in shaping microbiome diversity and biogeography, at the key scale of human communities.

**Keywords:**
Gut microbiome, social microbiome, Indonesia, hunter-gatherer, lifestyle, transition, sporulation

## Introduction

The gut microbiome is increasingly recognised as playing important roles in human health and disease (Valdes *et al.*, 2018; Sonnenburg and Sonnenburg, 2019). Yet, despite over a decade of sequencing effort, knowledge of extant diversity is limited as demonstrated by the vast numbers of novel species that continue to be discovered (Leviatan *et al.*, 2022; Carter *et al.*, 2023). Further, research is strongly focussed on urban, industrialised samples from a handful of over-represented countries (Abdill, Adamowicz and Blekhman, 2022), known to harbour limited microbiome variation with depletion of various phylogenetic groups (Yatsunenko *et al.*, 2012; Rosas-Plaza *et al.*, 2022; Carter *et al.*, 2023). As such, descriptions of the human microbiome, while increasingly extensive (Pasolli *et al.*, 2019; Almeida *et al.*, 2021; Gounot *et al.*, 2022), remain limited and exhibit significant regional and phylogenetic bias. The need to diversify sampling is clear and immediate (Abdill, Adamowicz and Blekhman, 2022).

Within this context, the Southeast Asian island country of Indonesia has the fourth largest population globally at over 270 million people and is among the most diverse regions worldwide, with over a thousand ethnic groups officially recognised across the archipelago (Na'im and Syaputra, 2012). Despite this, Indonesia is represented by only 0.01% of global metagenomic samples (Abdill, Adamowicz and Blekhman, 2022). About half of the population live in rural villages, though the country is experiencing rapid growth and urbanisation. The breadth of lifestyles represented, as well as the pace of change, represent clear opportunities to study the nature of lifestyle-associated microbiome transitions, and to broaden representation in microbial datasets.

The impact of lifestyle transitions on the microbiome has been studied in a range of global contexts. Broadly, urban populations show profoundly reduced microbiome diversity and divergent microbial composition e.g., an excess of *Bacteroides* relative to *Prevotella* and *Treponema*, compared to transitional and rural communities (Schnorr *et al.*, 2014; Obregon-Tito *et al.*, 2015; Gomez *et al.*, 2016). Certain hunter-gatherer groups are particularly diverse (Carter *et al.*, 2023), potentially reflecting the longer-term evolutionary context of the human microbiome with relevance to human health. Detailed work on migrant communities (Vangay et al. 2018) has been especially productive in revealing the timescales of microbiome transitions. However, transition is not a single axis, and associations between subsistence strategy and microbiome composition depend on various lifestyle factors and can vary between populations and studies (Gupta, Paul and Dutta, 2017; Rosas-Plaza *et al.*, 2022), interacting with numerous drivers to shape microbiome variation in complex ways.

Transitions ultimately arise from the behaviour of individuals living within communities, in the context of microbial diversity that may also be geographically structured (Suzuki *et al.*, 2022). At the individual level, environment is a key driver of microbiome composition (Rothschild *et al.*, 2018), along with diet (Zhernakova *et al.*, 2016; Wilson *et al.*, 2020), antibiotic use (Korpela *et al.*, 2016; Zhernakova *et al.*, 2016; Nagata *et al.*, 2022), birth practices (Penders *et al.*, 2006), and age (Yatsunenko *et al.*, 2012). Communities are increasingly a focus of research, with social interactions and transmission important drivers of strain diversity and sharing (Brito *et al.*, 2019; Valles-Colomer *et al.*, 2023) together shaping the 'social microbiome' (Debray, Tung and Archie, 2024), defined as the collective microbial metacommunity of a social group or network (Sarkar *et al.*, 2020, 2024). Yet the relationship between individual behaviour and community living – and indeed, how this interacts with the diversity of microbial physiologies and transmission routes – is hard to disentangle and only partially understood.

In this study, we build a region-specific gut microbiome database by sequencing faecal samples from 116 individuals from 8 communities (Supplementary Table S1). These communities occupy the islands of Bali, Borneo and New Guinea (Figure 1A), spanning an archipelago of approximately 2,500 km yet united by a tropical climate and interconnected recent and more ancient human histories. Critically, the groups also span a range of lifestyles, including transitional hunter-gatherers (Punan Batu), fisher-horticulturalists (Asmat), mixed forager-agriculturalists (Punan Aput and Punan Tubu-Hulu), horticulturalists (Basap), rural agriculturalists (Balinese in Pedawa), peri-urban residents (Punan Tubu-Respen) and fully urban residents (Balinese in Denpasar). Note that the Punan (or Penan) traditionally practised a hunter-gatherer lifestyle; the Punan groups in this study hence include several communities at different stages of integration with the market economy. The combination of geographic and lifestyle sampling with village-scale microbiome sampling, coupled with metagenome-resolved methods, enables us to identify and explore key drivers of local microbiome diversity. Finally, we describe the social biogeography of microbiome variation and novelty, identifying key bacterial traits associated with microbe geographic distribution and prevalence, with key implications for the understanding of global microbiome diversity.

## Results

### Novel Indonesian microbiome data significantly expands the number of genomes described in Southeast Asia

To characterise regional microbiome diversity, we first constructed a gut microbiome reference database based on faecal samples from our sample of 116 individuals, that were 250 bp paired-end sequenced (average yield = 10.63 Gb, Supplementary Table S2). We combined individual- and co-assembly methods to construct an assembly that was both sensitive to rare taxa and robust against assembly artefacts due to fine-scale genomic variation (Delgado and Andersson, 2022). This yielded 8,633 and 2,437 metagenome-assembled genomes (MAGs) respectively, and 11,070 MAGs total, hereafter referred to as the Indonesian Microbiome Ecology and Evolution (IndoMEE) reference database. This significantly expands the number of microbial genomes described in Southeast Asia (313 and 4497 Southeast Asian MAGs previously identified in Almeida *et al.* 2021 and Gounot *et al.* 2022 respectively). We combined and dereplicated the individual and co-assemblies to generate the final reference assembly, which comprised 1,304 (1,296 bacteria, 8 archaea) and 3,258 (3,247 bacteria, 11 archaea) species and subspecies level MAGs respectively. To classify MAGs, we used GDTB-tk to assign taxonomies based on the Genome Taxonomy Database (GTDBtk release 214). We identify a diverse set of bacterial species in IndoMEE, spread across 17 phyla, 23 classes, 50 orders, 111 families and 450 genera. Bacillota A is, by far, the most numerous phylum represented (~50% MAGs), followed by Bacteroidota, Bacillota and Pseudomonadota (Figure 1F).

### IndoMEE effectively captures regional diversity, and harbours many village-specific novel microbial species

To assess how well IndoMEE captures regional diversity, we evaluated and compared its mapping rate against state-of-the-art global and regional gut microbiome reference genome databases, namely the Unified Human Gastrointestinal Genome (UHGG) v2.0.1 and Singapore Platinum Metagenomes Project (SPMP) databases. Despite being assembled from a similar sample size (n = 116 for IndoMEE vs 109 for SPMP), IndoMEE significantly outperformed SPMP when applied on our Indonesian samples, mapping on average 72.7% of reads compared to 57.4% for the latter (Supplementary Figure S1). This highlights both the region-specificity of regional databases, and the higher diversity of microbiota captured in IndoMEE (based on diverse Indonesian villages; 1,304 species-level MAGs) compared to SPMP (urban Singapore; 685 species-level MAGs). IndoMEE also outperformed UHGG (67.8% mean mapping rate, Supplementary Figure S1), demonstrating that local sampling, capturing local diversity, can match or even exceed mapping rates from state-of-the-art global databases built from sample sizes several orders of magnitude larger. Merging of different assemblies led to marginal (i.e., sub-additive) increases in species-level MAGs and negligible improvement in mapping-rates (Supplementary Table S3, Supplementary Figure S1), implying that most MAGs are redundant (i.e., overlap between multiple datasets).

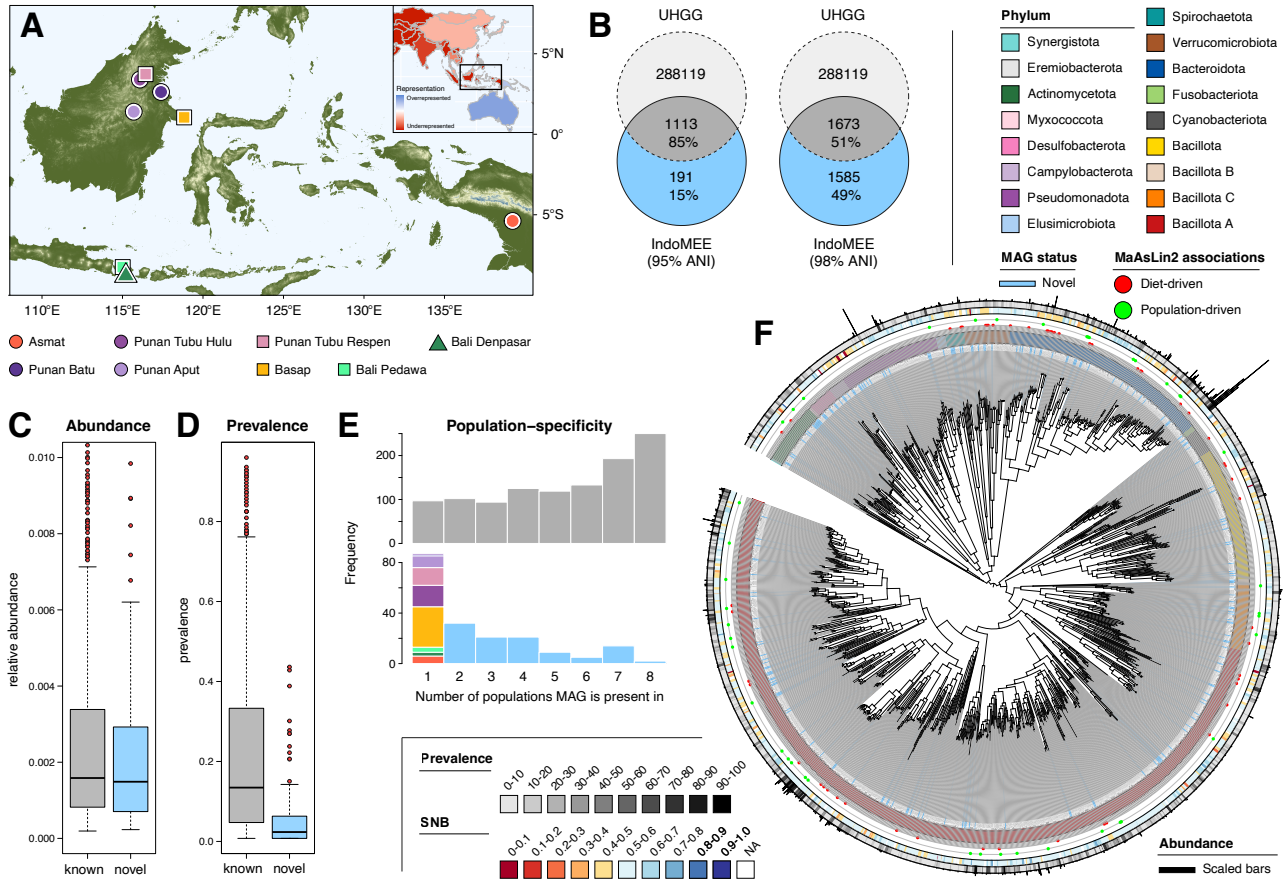


**Figure 1. Novel Indonesian microbiome database captures substantial microbial diversity and novelty.** **A**) Map of sampled Indonesian communities, with shapes denoting broad lifestyle categories (circles - remote; squares - rural; triangle - urban). Inset map shows countries' representation in current human microbiome datasets, relative to its share of the world population. Red and blue colours denote countries that are over-represented and under-represented relative to their population size, respectively. Countries with zero samples are indicated in dark red (modified from Abdill, Adamowicz and Blekhman, 2022). **B**) Numbers of novel (blue) and previously described (dark grey) MAGs in the newly constructed IndoMEE gut microbiome reference database, in comparison to the global Unified Human Gastrointestinal Genome (UHGG) database. **C**) Mean within-individual relative abundance, calculated across samples in which MAG is present, of novel MAGs versus known MAGs. No significant difference is observed between the two categories (Mann-Whitney two-sided test, $p = 0.15$). **D**) Prevalence of novel MAGs compared to known MAGs. Novel MAGs are significantly less prevalent (Mann-Whitney two-sided test, $p < 2.2 \times 10^{-16}$) than known MAGs. In box plots (**C** and **D**), centre lines denote medians, box edges indicate lower and upper quartiles, and whiskers extend to 1.5 times the interquartile range. **E**) Novel MAGs (bottom panel; coloured) are, relatively, more often unique to a single village than known MAGs (top panel; grey) in the IndoMEE database. Colours for village-specific novel MAGs correspond to that in panel A. **F**) Phylogenetic tree of IndoMEE MAGs annotated with phylum, status (novel or known), abundance, prevalence, social niche breadth (SNB) and if they are significantly associated with population or diet (inferred from MaAsLin2).

The large number of species- and subspecies-level MAGs recovered from a moderate sample size suggests that rural Indonesian villages are enriched for microbial diversity and novelty. To quantify this, we compared our newly assembled reference genomes against the largest available reference dataset (UHGG v2.0.1, full dataset: 289,232 MAGs). We discovered that 15% (191) and 49% (1,585) of our species- and subspecies-level MAGs, respectively, did not match the UHGG catalogue (Figure 1B), increasing the number of novel microbial species described from Southeast Asia ~2.7-fold (Almeida *et al.*, 2021; Gounot *et al.*, 2022). Among novel MAGs, several *Collinsella* species stand out as being both among the most prevalent (up to 43%) and most abundant (0.056%) (Supplementary Table S3). An undescribed *Prevotella* species has the overall highest abundance (0.13%) among novel MAGs and is also moderately prevalent (28%), while a species of the family Anaerovoracaceae is the most prevalent across novel MAGs (45%). While there is no significant difference in abundance overall between novel and known MAGs (Mann-Whitney two-sided test, $p = 0.15$; Figure 1C), novel MAGs are significantly less prevalent (Mann-Whitney two-sided test, $p < 2.2 \times 10^{-16}$; Figure 1D) and are more often found in a single community (Figure 1E, Supplementary Figure S1A) than known MAGs, whilst being distributed widely across the phylogenetic tree (Figure 1F). Further, novel MAGs are found in more similar microbial compositions (i.e., exhibit a lower social niche breadth score (von Meijenfeldt, Hogeweg and Dutilh, 2023)) than known MAGs, as measured by the average pairwise dissimilarity based on the inverse Spearman's rank correlation (Mann-Whitney two-sided test, $p = 6.8 \times 10^{-9}$; Supplementary Figure S1B-C). The village-specificity of novel MAGs emphasises the importance for local sampling to capture uncharacterised microbial diversity, which here tends to be rare and local. Overall, our novelty results demonstrate unique regional microbiome diversity in Indonesia, which is likely reflective of communities in the broader Southeast Asian region.

**Diverse lifestyle sampling reveals microbiome rural-to-urban transitions**

The human microbiome is known to diverge based on geography and lifestyle. In this context, IndoMEE is unique in sampling communities from widely separated tropical islands following diverse subsistence strategies and living conditions, which allows comparison of individuals across a broad socio-cultural and geographic range. We first sought to situate our data through comparison with a nearby, industrial outgroup: Chinese, Malay and Indian groups in Singapore, obtained from the SPMP dataset (Gounot *et al.*, 2022). To better investigate lifestyle variation, we further categorised the analysed communities into remote, rural, and urban groups based on observed parameters, including environment type, food sources, and remoteness (Figure 1A, Supplementary Table S1). We found that our Indonesian samples had higher diversity compared to SPMP samples, and there was a gradual and significant decline from remote to rural to urban communities (Figure 2A; all $p < 0.001$, pairwise Wilcoxon Rank-Sum tests). The decline was consistent across other alpha diversity metrics, including Faith's diversity index and raw species count (richness) (Supplementary Figure S2A). Within the urban group, the Denpasar group was significantly more diverse compared to Singaporeans (all $p < 0.001$, pairwise Wilcoxon Rank-Sum tests), demonstrating that city-living can be coincident with moderate microbial diversity.

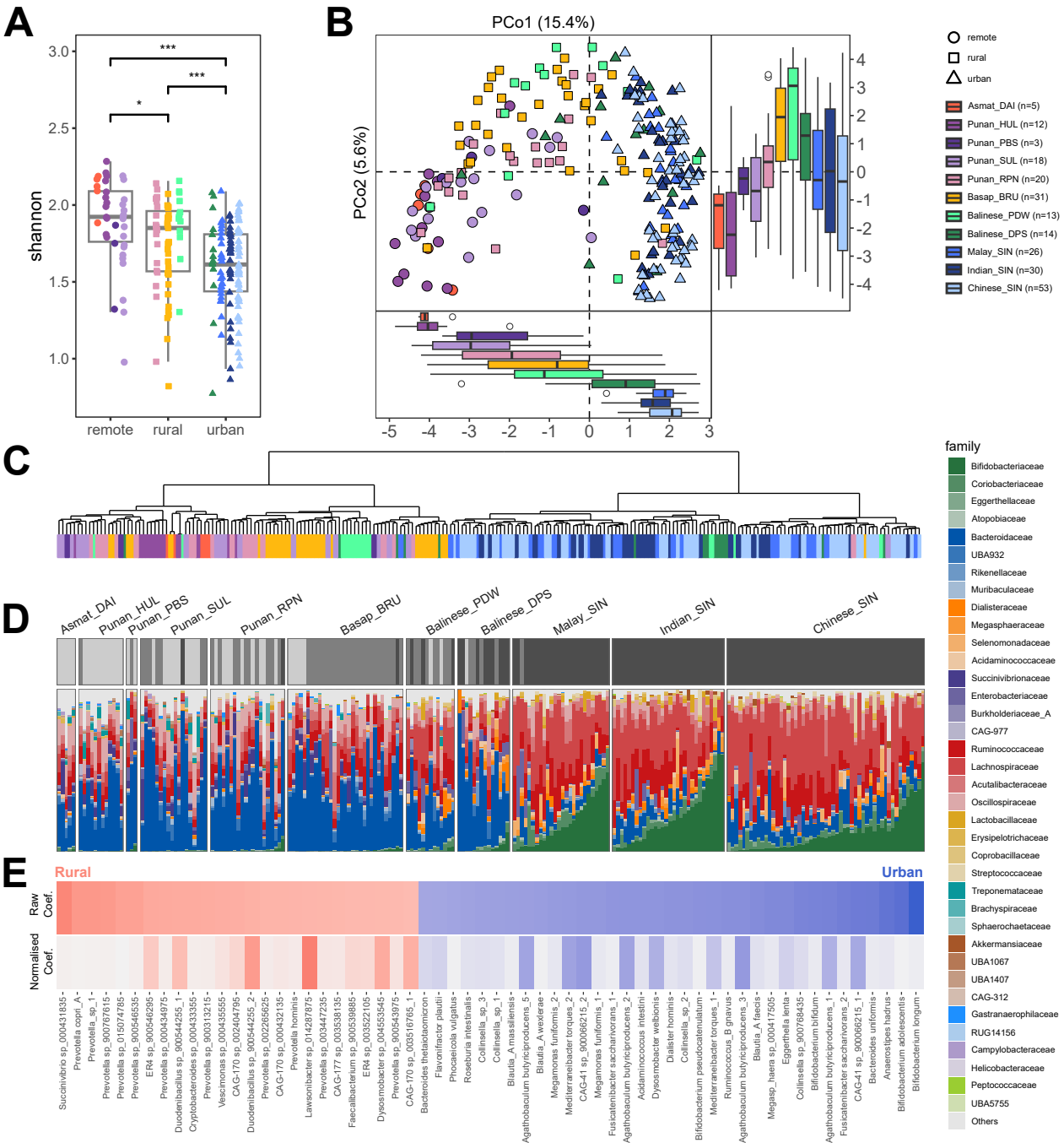


**Figure 2. Faecal microbiome diversity across diverse populations in Indonesia compared to Singapore.** **A**) Shannon's diversity index across 116 Indonesian and 109 Singaporean individuals across 11 distinct communities. Samples are grouped by lifestyle as remote, rural, and urban. *, ** and *** denote significance thresholds of $p < 0.05$, $p < 0.01$, and $p < 0.001$, respectively. **B)** Principal coordinate analysis of faecal microbiome based on Aitchison distance involving 1113 MAGs present in >2 individuals. Box and whiskers plot shows the positional distribution of individuals along the first and second principal axes. In box plots (**A** and **B**), centre lines denote medians, box edges indicate lower and upper quartiles, and whiskers extend to 1.5 times the interquartile range. **C)** Hierarchical clustering analysis of microbiome beta-diversity distance (Aitchison) group individuals into three clusters. Node-tips are coloured by population of origin (see legend). **D)** Relative abundances of bacterial families across 11 phyla and 11 communities. Families are ordered by abundance and only the top 4 most abundant families across phyla are shown. The panel above the bar plot shows three hierarchical sample clusters based on microbiome composition as depicted in **C** (light grey = remote; dark grey = rural; black = urban). **E)** Differential abundance analyses on 561 MAGs presents in >20 individuals in association with urbanisation (ordinal scoring: 0 = remote, 0.5 = rural, 1 = urban), adjusted for age and sex. Only the top 60 associated MAGs by absolute coefficient are shown. The top bar shows the raw coefficients and the bottom bar shows the coefficients normalised by the mean MAG abundance. Coefficient heatmap are coloured by association with urbanisation (red is rural, blue is urban).

We further analysed microbiome similarities between individuals using Principal Coordinate Analysis (PCoA). Individuals clustered by communities, but with a clear trend in variation that mirrors the lifestyle transition from remote to urban groups along the first principal axis (15.4% variance) (Figure 2B), coupled with divergence of rural villages on the second principal axis (5.6% variance). The association between individual microbiome composition and lifestyle groups (0 = remote, 0.5 = rural, 1 = urban) is highly significant (PERMANOVA, F = 30.54, all $p < 0.001$ obtained with permutations and adjusted for population group). These observations suggest a transitional state of the microbiome as communities become more urban. Hierarchical clustering indeed detects three primary sample clusters (Figure 2C), significantly associated with lifestyle groups ($\chi^2 = 225.8$; $p < 0.001$). The influence of lifestyle appears stronger than geographical proximity - evident in our data from Borneo and Bali. Both include communities that are regionally proximate, but with different lifestyles: Punan Tubu-Hulu (92% samples in remote Cluster 1; Figure 2C) is geographically close to Basap (81% samples in rural Cluster 2) in Borneo; and Pedawa (62% samples in rural Cluster 2) is close to Denpasar (64% samples in urban Cluster 3) in Bali. This result highlights the coherence of a transitional state in the microbiome and the importance of lifestyle over geography in structuring species-level microbiome diversity.

The majority of microbial families (76 of 113, 67%) are significantly different in abundance between sample clusters (Kruskal-Wallis test; $q < 0.05$). The greatest of these differences is the enrichment of Bifidobacteriaceae in urban Cluster 3 (43-fold enrichment over Clusters 1 and 2), accompanied by a depletion of Bacteroidales-UBA932 (17-fold), Succinivibrionaceae (62-fold), and Treponemataceae (1290-fold) (Figure 2D). To explore the difference at the species-level, we used MaAsLin2, focussing on 561 MAGs with prevalence > 20 individuals. Of these, 364 (67%) were found to be significantly associated with the urbanisation cline, emphasising the breadth and scale of microbiome response ($q < 0.05$) (Supplementary Table S4). Notably, three *Bifidobacterium* species (*B. longum. B. adolecentis,* and *B. pseudocatenulatum*) are among the most associated species with urbanisation (Figure 2E). *Bifidobacterium* enrichment was negatively correlated with *Treponema*, in accordance with previous work on rural and hunter-gatherer microbiomes (Schnorr *et al.*, 2014; Angelakis *et al.*, 2019; Stražar *et al.*, 2021; Rosas-Plaza *et al.*, 2022). We also observe replacement of *Prevotella (P. copri, Prevotella sp., P. hominis)* with *Bacteriodes/Phocaeicola* (*B. uniformis*, *B. thetaiotaomicron*, and *P. vulgatus*) with increasing urbanisation. Other notable enrichments in the urban populations are Bacillota A members: *Anaerostipes hadrus, Agathobaculum butyriciproducens, Blautia_A faecis, Blautia_A wexlerae*, and *Ruminococcus_B gnavus*; along with Actinomycetota members: *Eggerthella lenta* and *Collinsella sp.*; while Pseudomonadota members: *Succinivibrio sp.* and *Duodenibacillus sp.* were depleted. Indeed, clines in relative abundance over the lifestyle spectrum were observed for a wide range of genera (Figure S2B).

**Community, more than individual, lifestyle predicts microbiome differentiation in Indonesia**

To further investigate the drivers of this lifestyle signal, we analysed individual-level diet and daily activity data that was available for 122 Indonesian samples from 5 of the communities (Supplementary Figure S3A-B). Principal Component Analysis (PCA) of this expanded lifestyle data reveals substantial variation, highly structured by community, with PC1 (26% variance) driven by hunting as well as consumption of sago (palm) and pork (wild in remote/rural communities) in Asmat and Punan Tubu-Hulu (Figure 3A, Supplementary Figure S3C). These individuals contrast those from Pedawa and Denpasar, who had a higher frequency of watching TV and chicken consumption. On the other hand, PC2 (15.7% variance) was driven by higher consumption frequency of fish, egg and milk (Figure 3A, Supplementary Figure S3C). Further, we observed a strong linear correlation

between the first principal axes of our microbiome PCoA and lifestyle PCA ($R^2 = 0.562$, $p < 0.001$; Figure 3B), with substantial divergence of remote groups from others. The association was not significant if tested in a linear mixed model with population as a random effect (where population explained 54.5% of variance), suggesting that lifestyle differences between populations contribute very substantially to this relationship. Indeed, we do not observe correlations between diet and microbiome composition within individual groups (Supplementary Figure S3D), emphasising that the observed dietary associations primarily reflect variation between, rather than within, communities. Analysis at the village-scale corroborated this: a partial Mantel test of the correlation between village microbiome divergence and lifestyle while controlling for geographic distance was significant (Mantel test, R = 0.57, $p = 0.04$), while that between microbiome divergence and geographic distance controlling for lifestyle difference was not (Mantel test, R = 0.18, $p = 0.325$; see Methods, Supplementary Figure S3E).

We further explored whether the abundance of a single MAG could be predicted by specific diet components. We performed regression analyses using various univariate and multivariate combinations of food frequency data (7 variables: rice, sago, pork, egg, milk, fish, and chicken), as well as mixed-effects models with population as a random effect. Across 667 MAGs tested, univariate models predicting MAG based on single dietary factors revealed significant associations with sago, pork and chicken consumption, with a single MAG also associated with egg (Supplementary Table S4). Multivariate modelling using the three main dietary factors (sago, pork and chicken) identified 82 (12.3%) MAGs as significant (Figure 3C). Interestingly, comparison to a univariate model with population as the fixed effect revealed considerable overlap (73 of 82, 89.0%) as well as many more population associated MAGs (261 of 667, 39.1%; Figure 3D). Applying a mixed effect model incorporating the three dietary fixed effects and population as a random effect yielded 53 (7.9%) diet-associated MAGs (Figure 3E), including identifying two *Treponema_D succinifaciens* MAGs as sago-associated. As expected, community-level dietary trends (Supplementary Figure S3A) remain apparent in this analysis: pork-, sago- and chicken-associated MAGs tend to be enriched in Punan Tubu-Hulu, Asmat and Bali respectively. Separately, population also accounts for a considerable proportion of variance explained in the model (Figure 3F). MAGs most influenced by population are Prevotellaceae bacterium *UBA4372 sp.* (78.6%), Oscillospirales bacterium *UBA644 sp.* (65.4%), and *Vescimonas sp* (64.5%). Taken together, our analyses incorporating lifestyle data emphasise that host population is a stronger predictor of microbiome composition than individual diet, likely reflecting population-scale properties such as community lifestyle, environment or microbial transmission.

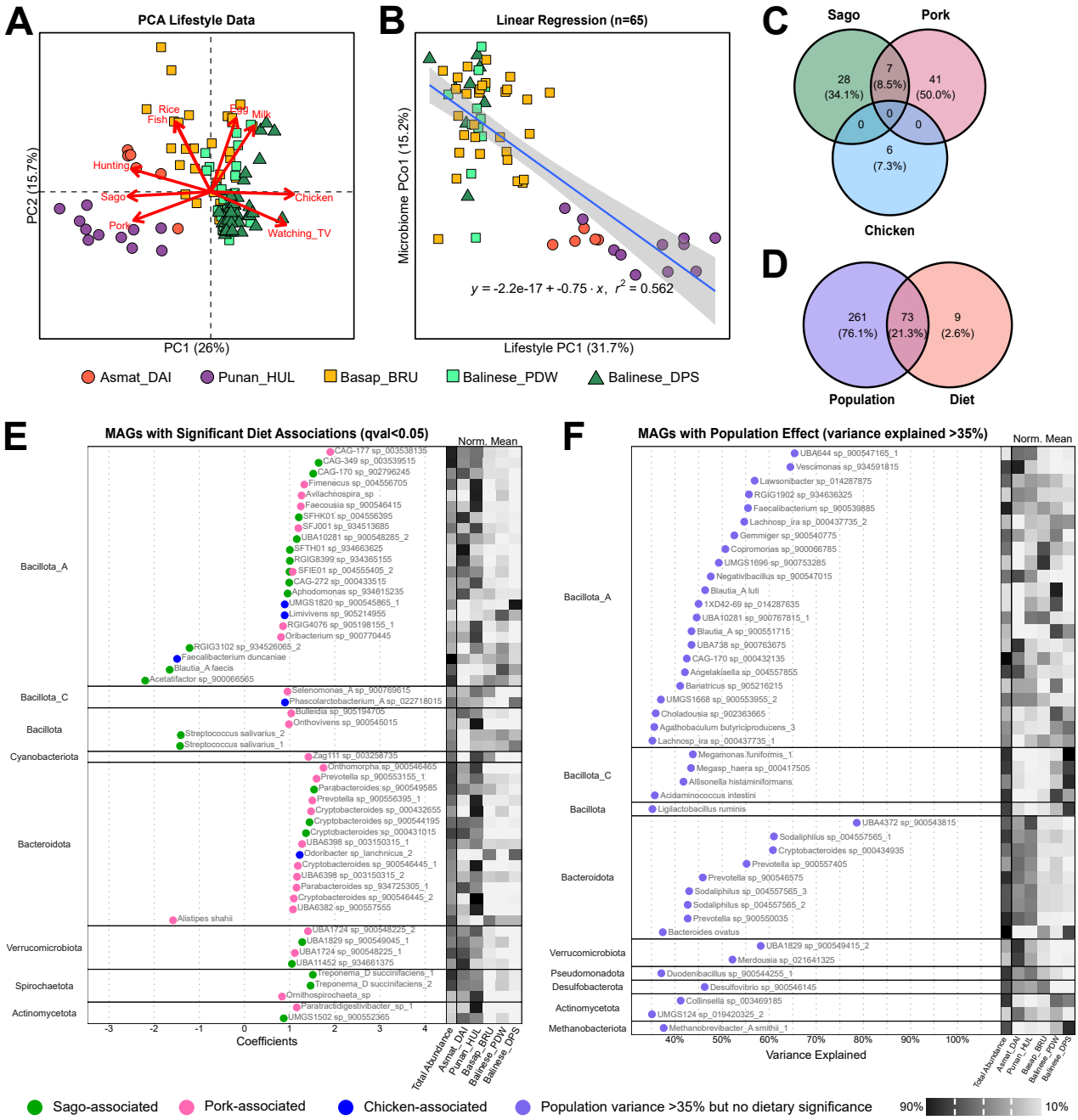


**Figure 3. Microbiome composition is associated with lifestyle factors in IndoMEE. A**) Principal Component Analysis (PCA) of lifestyle factors in 122 samples across 5 Indonesian populations. **B**) Linear regression analysis of the first principal axis (PC1) of the lifestyle PCA against PC1 of the microbiome PCoA (65 samples). **C**) Venn diagram of the count of MAGs associated with dietary factors identified in MaAsLin2 ($q$ < 0.05 is considered significant). **D**) Venn diagram of the total count of MAGs associated with population and diet as depicted by multivariate MaAsLin2 ($q$ < 0.05 is considered significant). **E**) MAGs significantly associated with diet ($q$ < 0.05), as identified with multivariate MaAsLin2 with population as a random effect. The right panel shows a heatmap of total and normalised mean abundance across 5 populations. **F)** MAGs with > 35% variance explained by population in the multivariate association analysis depicted in **E**. Villages are coloured as in Figure 1.

**Sporulation as a key bacterial trait that cascades up to the structure and novelty of the social microbiome**

The composition of the human gut microbiome in Indonesian populations is predicted in part by host lifestyle and community structure, but may also be driven by microbial properties that influence transmission and survival in non-host environments. We functionally annotated MAGs across a broad suite of functional classes including carbohydrate-active enzymes (CAZy), KEGG functional orthologs, COG clusters of orthologous genes, and several others (see Methods). We identified functions across the spectrum of the aforementioned classes, and a large number of uncharacterised (427,699) and novel (116,953) proteins, in our dataset (2,129,118 proteins total). Novel MAGs were enriched for both uncharacterised and novel proteins relative to known MAGs ($\chi^2$ tests, $p$= 7.0 x $10^{-16}$ and < 2.2 x $10^{-16}$ respectively). Focusing on KEGG functional orthologs (n = 6,450) and carbohydrate-active enzymes (n = 325), we performed PCA and find that clines in species composition (Figure 2B) are mirrored by functional structure (Supplementary Figure S4), with 22.1% and 21.8% variance explained respectively by PC1, and with consequent strong association with the lifestyle cline.

We then assessed which (KEGG) functions or genes were enriched or depleted in the novel MAGs relative to the known MAGs. Of 157 genes identified as significantly enriched or depleted, a significant proportion (17%; 27 genes) were identified as associated with sporulation (Figure 4A). Notably, all of these were inferred as significantly depleted in the novel MAGs. The transmission of microbes depends in part on their persistence in non-host environments, which can be facilitated by the development of a spore that exhibits minimal metabolic activity and high resistance to environmental stressors (Swick *et al*., 2016). However, a link to novelty has been less clear, though it may be hypothesised that this association (i.e., between novelty and sporulation) could be driven by lower prevalence (Browne *et al.*, 2021). To test if depletion in sporulation genes is associated with lower prevalence (and consequently MAG discovery), we regressed MAG sporulation gene counts $n_s$ (i.e., the number of distinct sporulation-associated genes present in a MAG; Supplementary Figure S5A) against MAG prevalence and the number of villages the MAG is present in. As bacterial spore-formation is a characteristic trait of Bacillota (previously Firmicutes) (Galperin, 2013), we assessed this both within and across Bacillota phyla, in addition to across the whole dataset (i.e., all MAGs across all phyla). Higher $n_s$ was significantly associated with higher prevalence in the whole, pan-Bacillota and Bacillota A datasets, while results were insignificant (or significant in only one of the prevalence metrics) for Bacillota B, Bacillota C and Bacillota ss (Figure 4B, Supplementary Figure S5). This may be partly attributed to the low sample size of these latter phyla compared to Bacillota A in our dataset (Figure 1F). Results across various annotations (e.g., KEGG and Pfam) were consistent, suggesting results are robust to annotation database (Supplementary Figure S5B-E). Additionally, we find that the master regulator gene essential for sporulation, Spo0A, is a strong predictor of $n_s$ ($p$ < 0.0001; Fig 4C), further supporting the link between sporulation and prevalence.

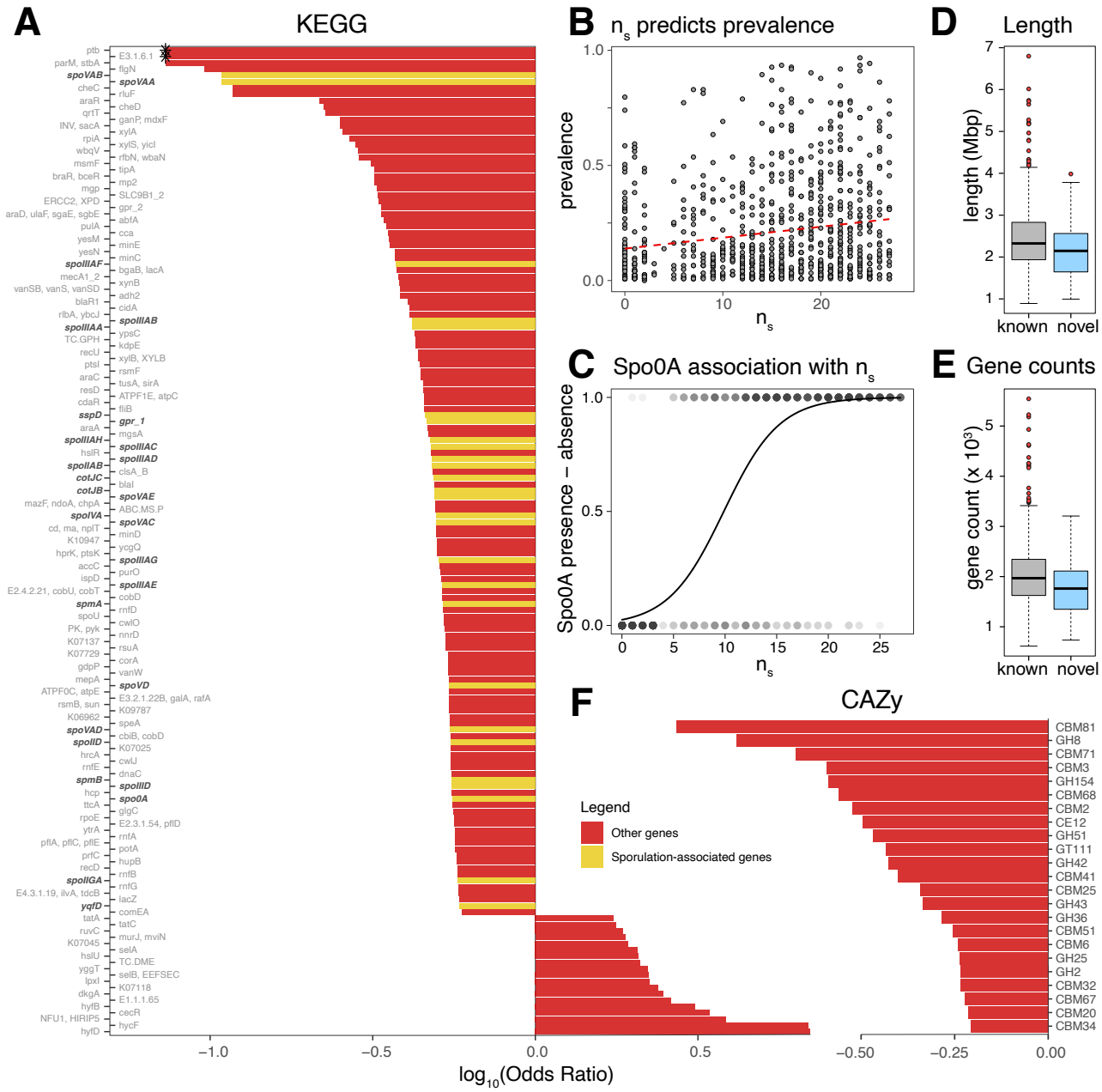


**Figure 4. Sporulation gene counts, which are depleted in novel MAGs, predict microbe prevalence. A)** Functional enrichment analysis (KEGG functional orthologs) of novel versus known MAGs, showing all significant enriched and depleted features. Sporulation-associated genes are coloured in yellow and annotated in bold. Significance and odds ratios were calculated via Fisher's exact test. Asterisks denote infinite odds ratios (i.e., zero denominator). **B)** MAG sporulation gene count ($n_s$) is associated with higher prevalence across Bacillota phyla (linear regression; $p < 5.7 \times 10^{-7}$). **C)** Logistic regression of MAG sporulation gene counts ($n_s$) against the presence-absence of the Stage 0 sporulation protein A (Spo0A) across Bacillota phyla ($p < 0.001$, Nagelkerke's pseudo-$R^2 = 0.805$). Circles with darker shades of grey reflect higher frequency of points. **D)** Length of novel MAGs compared to known MAGs. Novel MAGs are significantly smaller than known MAGs (Mann-Whitney one-sided test, $p = 0.001374$). **E)** Gene count of novel MAGs compared to known MAGs. Novel MAGs possess significantly fewer genes than known MAGs (Mann-Whitney one-sided test, $p = 0.003243$). In **D**) and **E**), centre lines denote medians, box edges indicate lower and upper quartiles, and whiskers extend to 1.5 times the interquartile range. **F)** Functional enrichment analysis of novel versus known MAGs for carbohydrate-active enzymes (CAZy), showing all significant features. Significance and odds ratios were calculated via Fisher's exact test.

Loss of sporulation capacity in Bacillota has been associated with greater host adaptation, and has been found to occur along with a suite of contemporaneous adaptations, including reduced genome size and narrowed metabolic capabilities (Browne *et al.*, 2021; Debray, Tung and Archie, 2024). Consistent with these expectations, we find that novel MAGs in IndoMEE are smaller (Mann-Whitney one-sided test, $p = 0.001374$), possess fewer genes (Mann-Whitney one-sided test, $p = 0.003243$) and exhibit reduced metabolic carbohydrate metabolism (Figure 4D-E) compared to known MAGs, implying that novel MAGs may be particularly enriched for host-adapted species.

Collectively, our results elucidate links between bacterial traits (i.e., sporulation capacity), degree of host adaptation, transmission and, ultimately, prevalence; with reduced prevalence consequently impacting the probability of previous sampling in existing databases and hence novelty in our analysis. Sporulation hence appears to be a key bacterial trait that cascades up to the structure and novelty of the social microbiome, and explains in part the village-level differentiation patterns identified across rural Indonesia.

## Discussion

Our work presents a new Indonesian gut microbiome dataset, IndoMEE, that incorporates diverse lifestyle sampling and substantially extends known MAG diversity of the Southeast Asian region. We find 15% of species-level MAGs and almost 50% of subspecies-level MAGs are novel, emphasising the limited representation of both non-industrial populations generally and Southeast Asia in existing databases. Intriguingly, many novel species have low-prevalence, often found in only a single village. These novel species show depletion of sporulation genes, and indeed we find a clear overall association between sporulation gene count and prevalence – demonstrating the impact of microbial traits cascading up, likely via transmission efficiency, to influence the distribution of taxa, and ultimately species discovery. The varied lifestyles represented by our sampling allowed us to study lifestyle-microbiome associations in detail. We identify key differences between Indonesian and highly urbanised Singaporean samples, and observe clines in abundance among a wide range of taxa. Within Indonesia, we find that an individual's community-membership is a better predictor of microbiome composition than their diet. While diet clearly has a pronounced impact on microbiome composition, our results suggest that, at least in our study region in Indonesia, this largely changes at the scale of communities; a key observation for understanding human gut microbiome variation.

Lifestyle transitions coincide with significant changes in diet, hygiene practices, healthcare (including antibiotic use), resource acquisition strategies, and human-environment interactions (Popkin, 2003; Reyes-García *et al.*, 2019), and have been associated with significant changes in the gut microbiome. For example, urban populations show a characteristic reduction in gut microbial richness compared to various agricultural and hunter-gatherer communities, in addition to compositional differences such as decreased *Prevotella* abundance (Mancabelli *et al.*, 2017; Vangay *et al.*, 2018; Rosas-Plaza *et al.*, 2022). How general or varied compositional trends are across different populations and lifestyles, however, remain unclear, particularly given continued limited representation of non-Western, non-industrialised, microbiome samples. Our work reveals intermediate (transitional) states in microbiome composition along our sampled lifestyle cline that spans urban, agricultural and hunter-gatherer contexts in Indonesia. Specific community comparisons are especially informative. For example, while urban Bali Denpasar clusters with Singapore, their microbiomes are also compositionally similar to more rural Indonesian groups; our samples from the long-term agricultural Bali Pedawa and recently transitioned Punan Tubu Respen

communities have similar composition despite different subsistence histories and disparate geographies (i.e., Bali and Borneo); and Punan Tubu Respen and Punan Tubu Hulu, who are from the same ethnic community but integrated into the market-economy to different extents over recent generations, have distinct microbiome compositions. Overall, our work complements studies of inter-generational shifts among migrants (Vangay *et al.*, 2018) and supports relatively rapid changes in community microbiomes - over the timescale of a few generations - in ways that are strongly predicted by lifestyle.

Our described lifestyle cline is further evident at the level of microbial taxa. For example, we find enrichment of *Bifidobacterium* species and reduction in *Treponema* species in urban groups, as has been previously reported (Nakayama *et al.*, 2015; Angelakis *et al.*, 2019; Stražar *et al.*, 2021; Gounot *et al.*, 2022). *Bifidobacterium* are prevalent in the human gut, particularly during infancy (Derrien *et al.*, 2022), and the association with urbanisation may reflect increased accessibility to wheat, dairy, and probiotic products (Lu *et al.*, 2023; Chonnacháin *et al.*, 2024), with potential immune response implications (Stražar *et al.*, 2021). We also identify a *Prevotella* and *Bacteroides* association with lifestyle, as has been previously described (Gorvitovskaia, Holmes and Huse, 2016). While *Bacteroides* and *Prevotella* both metabolise polysaccharides, they differ in bile tolerance, and so increased intake of animal fat and meat as associated with rural-to-urban dietary shifts is thought to favour *Bacteroides*. While hunter-gatherer and rural diets vary, enrichment of *Prevotella* has been observed in several hunter-gatherer communities (Rosas-Plaza *et al.*, 2022). We also identify a range of new lifestyle-associated taxa - indeed, more than half (67%) of MAGs tested were associated with the remote-rural-urban axis, with pronounced signals across a wide range of genera. Our study extends previous work on microbiome-lifestyle associations in Indonesia based on 16S data (Nakayama *et al.*, 2015; Febinia *et al.*, 2022), and highlights transitional states of the microbiome in a rural-to-urban lifestyle cline.

Major shifts in the microbiome can occur rapidly at the individual scale (David *et al.*, 2014), however, the nature of lifestyle-associated microbiome transitions in human communities is less clear. Our results demonstrate that lifestyle associations are driven more by between-community differences than by within-community differences. Furthermore, at the scale of microbial species, population is far more frequently identified as a predictor of relative abundance (76.1% of associations) than dietary factors alone (2.6%). Diet remains a fundamental driver of microbiome composition, but may explain a minor part of compositional variance (5-20%) as noted across a range of study designs (Johnson *et al.*, 2020). Our results thus emphasise the importance of population-level processes – e.g., shared lifestyle, environments and more active processes like food sharing and transmission (Debray, Tung and Archie, 2024; Valles-Colomer *et al.*, 2023) – in shaping the social microbiome of human communities.

Beyond lifestyle, our work points towards the role of microbial traits, specifically sporulation, in revealing host-microbe dependence and in shaping microbiome structure. In our study, novel MAGs are significantly depleted for sporulation genes. Sporulation in bacteria is exclusive to Bacillota (previously Firmicutes) (Galperin, 2013), where sporulation is considered an ancestral trait. A loss of sporulation capacity has been linked to greater host adaptation, as it mandates more direct modes of between-host transmission to avoid non-host environmental exposure (Browne *et al.*, 2021; Debray, Tung and Archie, 2024). The resultant, altered transmission cycle and greater reliance on host are hence expected to drive a suite of contemporaneous adaptations, including reduced genome size and narrowed metabolic capabilities (Browne *et al.*, 2021). Our novel MAGs exhibit these tell-tale signatures of host dependence, suggesting that they may be particularly enriched for host-adapted species. Furthermore, we reveal that microbe's sporulation capacity correlates

significantly with microbe prevalence. While a link between sporulation, transmission and prevalence has been proposed (Browne *et al.*, 2021), the broader implications of this for characterising microbiome diversity and particularly MAG discovery has remained unclear. Our results draw a direct link between microbial physiology, prevalence, the social microbiome, and MAG discovery.

Our diverse gut microbiome sampling from Indonesia substantially extends known genome diversity in Southeast Asia, and identifies clinal patterns of microbiome variation that mirror community-scale lifestyle trends. By combining analyses on novelty and function with prevalence and population structure, our work further highlights a conceptual bridge between bacterial physiology, prevalence, community divergence and species discovery. Further sampling efforts in under-represented regions like Indonesia, particularly at the village scale and incorporating diverse lifestyles, will likely discover many more low-prevalence, often non-sporulating and host-adapted species, and offer further insights into the processes structuring microbiome composition at local and global scales.

**List of supplemental information**

Supplementary Information document. Figures S1–S5 and Table S1.
Table S2. Excel file containing additional data too large to fit in a PDF, related to Figures 1-3.
Table S3. Excel file containing additional data too large to fit in a PDF, related to Figure 1.
Table S4. Excel file containing additional data too large to fit in a PDF, related to Figures 2-3.

**Data and code availability**

Raw metagenomic sequencing reads for the 116 faecal samples will be deposited and made available at the European Genome-phenome Archive (EGA). Code used in this study is available at the GitHub repository: https://github.com/hirzi/IndoMEE.

**Acknowledgements**

The authors thank communities and volunteers involved in this research, as well as Prof. J Stephen Lansing, A/Prof. Andrew J. Holmes, Prof. I Wayan Weta, Datuk Karim, Ahmad Arif, members of the Genome Diversity and Diseases Laboratory field team (Prisca C. Limardi, Isabella Apriyana, Gludhug A. Purnomo, Bertha L. Utami, and Leonard Taufik), and university students at the Faculty of Medicine of Udayana University for supporting and assisting the community engagement and sample collection of this study between 2015 and 2021. We are grateful to local officials in Tanjung Selor and Malinau, and the directors of the Eijkman Institute for Molecular Biology (Prof. Sangkot Marzuki and Prof. Amin Soebandrio) for facilitating this study. This research was funded by the European Union's Horizon 2020 research and innovation programme under grant agreement no. 950610 and an NTU Presidential Postdoctoral Fellowship to G.S.J. (2018-2020), a Wellcome Trust International Training Fellowship (no. 222992/Z/21/Z) to P.K., a block grant from the government of the Republic of Indonesia to the Eijkman Institute for Molecular Biology through the Ministry of Research and Technology/National Agency for Research and Innovation (IMELDA Project, 2014-2021), the International Program Development Fund to A/Prof. Andrew J. Holmes, H.S., and S.G.M. (2013), Australia's Department of Foreign Affairs and Trade through the Australia Awards Scholarship Fieldwork Entitlement to C.A.F. (2014-2016), and the Cambridge Trust Scholarship to C.A.F. (2022-2024).

**Author's contributions**

C.A.F., H.L., S.G.M, and G.S.J. conceptualised the study. C.A.F., P.K., L.P., D.W., G.I., H.S. and S.G.M. facilitated and performed sample and data collection. C.A.F., H.L., P.K. and L.P. performed data curation. C.A.F., P.K., L.P., S.G.M., and G.S.J. developed field and laboratory protocol. C.A.F.,

H.L., A.A., and G.S.J. developed bioinformatic and statistical analysis pipelines. H.L. performed metagenomic assembly and annotation; and C.A.F., H.L., and J.L. performed data analyses. C.A.F., H.L., and G.S.J. wrote the manuscript, which all authors revised. G.S.J. and S.G.M. administered and supervised the study.

**Declaration of interests**
The authors declare no competing interests.

## Methods

**Study enrolment, data, and sample collection**
Community engagement, official permissions, and sample collection started in 2015 and was completed in 2021. Prior to collecting the samples, we liaised with local officials and leaders (e.g., village elders) to explain and seek permission to perform the study. During enrolment, we collected written informed consent from the volunteers. Data and sample collection were done in one instance (cross-section) in each community on separate occasions: Denpasar in 2015, Respen Tubu in 2018, Long Sule in 2018, Teluk Sumbang 2019, Punan Batu in 2019, Asmat in 2019, Rian Tubu in 2020, and Pedawa in 2021. We only included adults in this study (> 18 years old).

The study was approved by the Eijkman Institute for Molecular Biology Research Ethics Commission (project No. 80/2014, No. 119/2019, and No. 126/2019), Udayana University Faculty of Medicine and Sanglah Hospital Ethic Commission (No. 1286/UN.14.2/Litbang/2014 and No. 1742/UN14.2.2.VII.14/LT/2021), the Mochtar Riady Institute for Nanotechnology Ethics Committee (006/MRIN-EC/ECL/III/2022) and the University of Cambridge Human Biology Ethics Committee (HBREC.2021.01).

We collected participant metadata including age and family tree. We assessed dietary intake and daily activities (i.e., hunting, farming, walking) through interviews. The diet interview was conducted using a semi-quantitative food questionnaire, in which participants reported the number of servings of specific food items (e.g., white rice) consumed and the frequency (daily, weekly, monthly, or yearly). The questionnaire included a list of commonly consumed foods including staple foods (e.g., rice, sago, tubers, potatoes, and noodles), meat products (e.g., chicken, pork, fish, and beef), legumes, leafy vegetables, eggs, dairy products, and other market products (e.g., snacks and confectioneries).

The physical activity data captured the intensity of routine activities by measuring their frequency (times per day, week or month) and duration (in hours or distance). Examples included farming, hunting, foraging, fishing, and commuting (whether on foot or by other means). Questions on sedentary behaviours such as sleeping, watching TV, and casual socialising were also included. Use of vehicles was recorded if relevant.

In total, we interviewed 172 individuals for lifestyle data and obtained complete dietary data from 122 individuals, of which 110 were lean and 12 were obese (body mass index > 30). Faecal samples for the gut microbiome analysis were obtained from 116 individuals across the eight communities, of which 65 had fully completed the lifestyle questionnaire. For Bali (Pedawa and Denpasar), we stored faecal samples in −20 °C within 4 hours of defaecation and transferred them to −80 °C once they arrived in the lab. For other communities, samples were preserved at room temperature using

OmnigeneGUT (DNA Genotek) for a maximum of 60 days, adhering to the manufacturer's guidelines.

Singaporean metagenomic shotgun sequence data were obtained from the public repository of the Singapore Platinum Metagenomes Project (NCBI project accession number PRJEB49168) (Gounot *et al.*, 2022). The participants for this study were recruited cross-sectionally from a multi-ethnic cohort, comprising 109 subjects: 65 males and 44 females, between the ages of 48 to 76 years old.

**Lifestyle categories**

In this study, we categorised the sampled communities into remote, rural, and urban groups to facilitate comparisons among populations with shared traits. This categorisation is based on observed parameters, including environment type, food sources, and remoteness (detailed in Supplementary Table S1). Specifically, Asmat, Punan Batu, Punan Tubu Hulu, and Punan Aput are classified as remote; Punan Tubu Respen, Basap, and Balinese Pedawa as rural; and Balinese Denpasar and the Singaporean samples from the SPMP project as urban. These classifications are not intended as official designations, acknowledging the variability in definitions across regions and countries, and the interplay of factors such as population density, proximity to economic centres, infrastructure, economic diversity, and development across sectors.

**DNA extraction and sequencing**

DNA extraction was performed within three weeks of sample collection to minimise storage time. Extraction was conducted in a Level-2 Biosafety Cabinet using the PowerFecal DNA Isolation Kit (MO BIO, Cat# 12830-50) for the Denpasar samples. For the remaining samples, the QIAamp PowerFecal DNA Kit (QIAGEN, Cat# 2830-50), the rebranded version of the original kit, was used. Extractions used 0.25 g of stool for each sample and included extraction blanks (250 µL elution buffer 10 mM Tris, Qiagen Cat# 19086) as a negative control. For the Denpasar samples, homogenisation was performed using a Vortex Genie II (Sigma-Aldrich, Cat# Z258423) at maximum speed for 15 minutes (Febinia *et al.*, 2022). The remaining samples were homogenised using the FastPrep-24 system (MP Biomedicals, Cat#116004500) at 5.5 m/s for two cycles of 30 seconds each, with a 5-minute interval between cycles. The quantity of extracted DNA was measured using the Qubit dsDNA Broad Range Kit (ThermoFisher Scientific, Cat# Q32853). DNA purity was assessed with a Nanodrop Spectrophotometer ND-1000 (ThermoScientific), and genome integrity was evaluated via gel electrophoresis in 1% agarose using 1× TBE buffer. Samples with contamination or low DNA quantity underwent a second extraction. If the quality improved, the second extracts were used for further analysis.

We sequenced a total of 116 samples in total, along with 8 technical replicates, 2 positive controls (ZymoBIOMICS Microbial Community DNA Standard by Zymo Research Cat# D6305), and 4 negative controls (for sampling: Blank, for extraction: Buffer EB by Qiagen Cat# 19086, for library prep: Nuclease Free Water by Ambion-Invitrogen Cat# AM9937). DNA libraries were prepared using Illumina's TruSeq Nano DNA Kit. Sequencing was performed on the Illumina NovaSeq 6000 platform to generate 250 bp paired-end reads, with an average yield of 10.63 Gb / sample. Following quality control, we acquired high-quality faecal metagenomic sequences from 116 samples for subsequent analyses (Supplementary Table S2).

**Metagenomic assembly**

We assembled metagenome-assembled genomes (MAGs) using a modified version of the metaWRAP pipeline (version 1.3.2) (Uritskiy, DiRuggiero and Taylor, 2018), from the sequence data of 116 samples. Reads were first trimmed for adapters and by quality using Trim Galore (version

0.6.10) (Krueger, 2016), under default settings (-q 20 --length 20 -e 0.1). Trimmed reads were then aligned to the complete human (host) reference genome (CHM13v2.0) via bmtagger (version 1.1.0), to remove human reads from the metagenomic sequences.

Host-cleaned metagenomic reads were then assembled via a hybrid approach that combined individual and co-assembly methods. Combining reads via co-assembly increases read depth which facilitates the recovery of species that may be too rare to detect in individual samples, however, mixing data from closely related taxa (strains) can adversely affect the assembly process, as intraspecific genomic variation can cause de Bruijn graphs to incorporate numerous alternative paths (Delgado & Andersson 2022). Reads were assembled using MEGAHIT (version 1.2.9) for each individual (individual assembly) and for each population (co-assembly). Assembled contigs were then binned via MaxBin2 (version 2.2.4), metaBAT2 (version 2.12.1) and CONCOCT (version 1.0.0) (individual assembly) or MaxBin2 exclusively (co-assembly). When more than one binning software was used (individual assembly), we used metaWRAP's bin-refinement module to produce a consolidated, improved MAG set, based on MAG completeness and contamination rates as estimated from CheckM (version 1.0.18) (Chklovski *et al.*, 2023). Specifically, MAGs were scored based on the scoring function $S = Completeness - 5 * Contamination$. We additionally applied a hard filter, requiring a minimum completeness of 50% and allowing a maximum contamination of 5% for MAGs. This yielded 8,633 and 2,437 MAGs for the individual assembly and co-assembly respectively. The resultant set of MAGs are non-redundant and comprise many replicates across individuals and across populations. To get representatives at species and subspecies levels, MAGs were dereplicated using dRep (version 3.4.5) (Olm *et al.*, 2017) at 95% and 98% average nucleotide identity (ANI) respectively, using the fastani algorithm (Jain *et al.*, 2018), CheckM2 (version 1.0.2) (Chklovski *et al.*, 2023) estimates of completeness and contamination, and the following dRep parameters: -pa 0.9 -sa 0.95 -nc 0.3 -l 50000 -comp 50 -con 5. Dereplicated MAGs were subsequently filtered for $S \geq 50\%$ and for chimerism and contamination content via GUNC (version 1.0.6) (Orakov *et al.*, 2021). For the latter, we filtered out genomes that matched all the following criteria: i) flagged by GUNC (i.e., clade_separation_score > 0.45; contamination_portion > 0.05; reference_representation_score > 0.5), ii) are singletons (dRep clusters with only one member) and iii) < 90% completeness based on CheckM2. Finally, we combined and dereplicated the resultant individual assembly and co-assembly, using the same dRep parameters as above, to generate the final assembly.

**Taxonomic assignment and phylogenetic tree construction**

To taxonomically classify MAGs, we used GTDB-tk (version v2.3.2) (classify_wf) to assign taxonomies to MAGs based on the Genome Taxonomy Database (GTDB release 214), and to build multiple sequence alignments for constructing phylogenetic trees. We generated phylogenetic trees via IQ-TREE (version 2.3.3) (Nguyen *et al.*, 2015) with the ModelFinder Plus (-m MFP) argument, which automatically performs model selection and subsequently constructs trees using the inferred best-fit model. Best-fit models according to the Bayesian information criterion (BIC) for bacteria and archaea were LG+R10 and LG+F+G4, respectively. We then used the Interactive Tree Of Life (iTOL) (version 6) online plotting tool to annotate and plot phylogenetic trees.

**Quantifying abundance, prevalence and mapping rates**

To evaluate mapping rates across different assemblies, and calculate MAG relative abundances, we used metamap (Almeida, 2021c). metamap maps paired-end metagenomic reads to a reference genome using bwa-mem (version 0.7.17-r1188) (Li, 2013; Jung and Han, 2022), and filters mapped reads based on nucleotide identity, mapping score and alignment fraction using SAMtools (version 1.9) (Danecek *et al.*, 2021). For each MAG in the reference, read counts, depth and (observed and

expected) breadth of coverage are then calculated; with expected breadth given as $breadth_{exp} = 1 - e^{-0.883*coverage}$ (Olm *et al.*, 2021). Because of mis-mapping, read counts greater than zero do not necessarily imply that a MAG is present in the sample (Olm *et al.*, 2021). To consider a MAG as present and minimise false positives, we required mapped reads to cover ≥ 50% of sites in a MAG (observed breadth ≥ 50%), or observed breadth ≥ 25% and ratio of observed breadth:expected breadth ≥ 0.8 (Olm *et al.*, 2021). We alternatively tested a more permissible filter (observed breadth ≥ 5% breadth and observed breadth:expected breadth ≥ 0.3).

To calculate relative abundance from raw read counts, we first divided the number of reads mapped to a MAG by the total number of reads in the metagenomic sample, and then normalised this by the length of the MAG (to control for the varying lengths of MAGs). We then divided the resultant normalised read count of each MAG by the sum of normalised counts in the metagenomic sample to get the relative abundance. MAG prevalence was calculated by enumerating the number of samples a MAG was present in, and subsequently dividing by the total sample size. MAG social niche breadth (von Meijenfeldt, Hogeweg and Dutilh, 2023) was calculated from the relative abundance table, with mean pairwise dissimilarity based on the inverse Spearman's rank correlation of species (MAGs).

For parity, all databases were dereplicated at 95% ANI (i.e., at species-level representation) when comparing mapping rates across databases. To assess redundancy and overlap between UHGG (4,744 species-level MAGs), IndoMEE (1,304 species-level MAGs) and SPMP (646 species-level MAGs) databases, we combined and dereplicated these databases at 95% ANI to construct merged reference databases. This yielded 4,860 and 4,941 species-level MAGs for IndoMEE+UHGG and IndoMEE+UHGG+SPMP databases respectively — resulting in marginal (sub-additive) additions in species-level MAGs and suggesting considerable (though not total) redundancy with UHGG.

**Profiling novelty**

To identify novel species and subspecies-level taxa among our assembled MAGs, we queried our MAGs against the largest available reference dataset (UHGG v2.0.1, full dataset: 289,232 MAGs) using magscreen (Almeida, 2021b). Magscreen first calculates Mash distances (Ondov *et al.*, 2016) between query MAGs and reference MAGs (i.e., in the reference database) to identify the closest reference MAG (to the query MAG) and, based on this and the alignment fraction as calculated using MUMmer, calculates the average nucleotide identity (ANI) of the query MAG and its nearest reference MAG. Following dereplication (dRep; -pa 0.9 -sa 0.95 -nc 0.30 -cm larger -comp 50 -con 5) and GUNC based quality-control filtering (same parameters as detailed in Metagenomic assembly), a MAG is identified as novel at the species level if the ANI to the closest representative in the reference is less than 95%, and at the subspecies level if the corresponding ANI is less than 98%.

**Microbiome composition analysis**

To investigate patterns in microbiome diversity across communities, we employed the microbiome relative abundance profiles of 116 Indonesian with 109 Singapore individuals, the latter derived from data obtained from a previous study (Singapore Platinum Metagenomes Project/SPMP) (Gounot *et al.*, 2022). The Indonesian samples comprise members from i) remote communities: Asmat_DAI (5 samples), Punan_PBS (3 samples), Punan_HUL (12 samples), and Punan_SUL (18 samples); ii) rural communities: Punan_RPN (20 samples), Basap_BRU (31 samples), and Balinese_PDW (13 samples); and iii) an urban community: Balinese_DPS (14 samples). The urban Singaporean samples consist of 3 groups: Chinese_SIN (53 samples), Malay_SIN (26 samples), and Indian_SIN (30 samples). Diversity analysis and statistical tests were performed using R (v.4.4.0). Alpha and

beta diversity were calculated from the relative abundance matrix using abdiv v.0.2.0 (Bittinger, 2020) and vegan v.2.6-4 (Oksanen *et al.*, 2022) packages, respectively. To characterise overall microbiome diversity, we calculated species richness using observed MAG count and Faith's PD, and species diversity using Shannon's Index, for each sample. To assess between-sample divergence, we calculated microbiome dissimilarity (beta diversity) using Aitchison distance (Aitchison *et al.*, 2000), and transformed and visualised the results using Principal Coordinate Analysis (PCoA). Sample hierarchical clustering analysis employed Ward's method (Murtagh and Legendre, 2014).

**Microbiome association with lifestyle and diet**

Lifestyle associations were examined among non-obese individuals, comprising 108 Indonesian and 109 Singaporean samples. To investigate differences in MAG abundance across remote, rural, and urban communities, we conducted differential abundance analyses using MaAsLin2 (Mallick *et al.*, 2021). For this analysis, we required MAGs to be present in > 20 individuals (542 of the total 1,304 MAGs), and modelled community type as an ordinal variable to reflect a linear gradient of lifestyle across populations (0 = remote, 0.5 = rural, 1 = urban).

To further assess the influence of lifestyle on the microbiome in Indonesian communities, we analysed diet and activity data. Lifestyle and activity-based distances between individuals were visualised using Principal Component Analysis (PCA) in a subset of 122 individuals (including 12 with obesity) across five communities: Asmat_DAI, Punan_HUL, Basap_BRU, Balinese_PDW, Balinese_DPS. For microbiome-lifestyle associations, we focused on 65 non-obese individuals with complete data (49 individuals were excluded due to the absence of faecal samples, and 8 due to obesity). We performed linear regression to assess the association between lifestyle factors (from the PCA on questionnaire data) and microbiome composition (from the PCoA on microbiome Aitchison distance). To further evaluate population-specific associations, we performed separate linear regression analyses for each population.

To test the effect of geography and lifestyle on microbiome divergence, we first calculated population mean relative abundances for MAGs and population averages for lifestyle data. Pairwise distances were then calculated in terms of Aitchison distance for the former and $\chi^2$ distance for the latter.
Great-circle geographic distances between populations were measured using coordinates approximating the central location of each population. Total-permutation partial Mantel tests were then performed to assess the correlation between microbiome divergence and geographic distance accounting for lifestyle distance (i.e., *microbiome divergence ~ geographic distance | lifestyle distance*), and between microbiome divergence and lifestyle distance accounting for geographic distance (*microbiome divergence ~ lifestyle distance | geographic distance*).

To disentangle MAG correlations with lifestyle and population (in 73 individuals with complete data), we used MaAsLin2, initially performing univariate regressions of MAG abundances (prevalence filter > 5 samples) with each diet factor and population group separately. We counted the total number of associated MAGs in each case and compared the collective number of significant associations in any dietary category (diet-associated MAGs; only chicken, sago and pork had significant associations) to any population (population-associated MAGs). This analysis hence captures to what extent diet or population might be used to predict MAG abundances, and how often these signals overlap. To better distinguish between population and dietary drivers, we performed further mixed effects models with population as a random effect and the three previously-identified dietary factors (chicken, sago and pork) as fixed effects. We calculated the variance explained by the

population random effect to identify population-driven MAGs, and considered MAGs with a significant association with any of the three dietary factors as diet-driven.**Functional annotation**
To functionally annotate the IndoMEE reference database, we used genofan (Almeida, 2021a). Genofan first calls prokka (version 1.14.6) (Seemann, 2014) to identify coding regions and annotate protein-coding genes, before annotating identified genes across a broad suite of functional classes via AMRFinder (version 3.11.14) (antimicrobial resistance), gutSMASH (version 1.0.0) (primary metabolic gene pathways), antiSMASH (version 6.0.1) (secondary metabolite biosynthetic gene clusters), dbCAN2 (version 2.0.11) (carbohydrate-active enzyme and substrates (CAZy)), KOFams (HMMER version 3.4) (KEGG (Kyoto Encyclopedia of Genes and Genomes) functional orthologs) and eggNOG-mapper (version 2.1.3) (GO (gene ontologies), COG (clusters of orthologous genes), Pfam (protein families), etc.). Genofan and constituent tools were run under default settings.

**Functional enrichment analysis**
To identify which functions or genes were significantly enriched or depleted in the novel MAGs relative to the known MAGs, we performed Fisher's exact test on gutSMASH, antiSMASH, CAZy, AMRFinder, Pfam, GO and KEGG annotated features. These functional features were considered as present or absent and tabulated in 2 x 2 contingency tables of (novel, known):(present, absent), for each functional feature or gene. For COG, functional enrichment was assessed using Wilcoxon rank-sum test (two-sample, unpaired), given that COG categories are broader and return numerous hits per MAG; and hence better considered as counts data. For both sets of tests, we filtered out MAGs below 90% completeness, given that varying rates of completeness may bias enrichment results. For tests based on feature counts (i.e., COG), we normalised counts by MAG length (total feature counts), to control for varying MAG lengths. Test-inferred significance was adjusted for False Discovery Rate (FDR) and final significance was taken as that with FDR-adjusted $p < 0.05$.

**Functional composition and structure**
To visualise compositional patterns in function and carbohydrate-active enzymes across samples, we calculated KEGG and CAZy (gene) counts for each MAG. Gene counts were weighted by the abundance of the MAG in the sample (individual). A centred log-ratio (CLR) transformation was applied to this table of abundance-weighted gene counts, to map compositional (i.e., Aitchison simplex space to Euclidean space), before applying an unscaled PCA. Mathematically, this is equivalent to applying a principal coordinate analysis on the Aitchison distance matrix.

## References


Abdill, R.J., Adamowicz, E.M. and Blekhman, R. (2022) 'Public human microbiome data are dominated by highly developed countries', PLOS Biology, 20(2), p. e3001536. Available at: https://doi.org/10.1371/journal.pbio.3001536.

Abecasis, A.B. et al. (2013) 'A Genomic Signature and the Identification of New Sporulation Genes', Journal of Bacteriology, 195(9), pp. 2101–2115. Available at: https://doi.org/10.1128/jb.02110-12.

Aitchison, J. et al. (2000) 'Logratio Analysis and Compositional Distance', Mathematical Geology, 32(3), pp. 271–275. Available at: https://doi.org/10.1023/A:1007529726302.

Almeida, A. et al. (2021) 'A unified catalog of 204,938 reference genomes from the human gut microbiome', Nature Biotechnology, 39(1), pp. 105–114. Available at: https://doi.org/10.1038/s41587-020-0603-3.

Almeida, A. (2021a) 'GenoFan - Genome functional annotation pipeline'. Available at: https://github.com/alexmsalmeida/genofan (Accessed: 20 November 2024).

Almeida, A. (2021b) 'MAGscreen - discovering new microbial species'. Available at: https://github.com/alexmsalmeida/magscreen (Accessed: 20 November 2024).

Almeida, A. (2021c) 'metaMap - quantifying genomes in metagenomes'. Available at: https://github.com/alexmsalmeida/metamap (Accessed: 20 November 2024).

Angelakis, E. et al. (2019) 'Treponema species enrich the gut microbiota of traditional rural populations but are absent from urban individuals', New Microbes and New Infections, 27, pp. 14–21. Available at: https://doi.org/10.1016/j.nmni.2018.10.009.

Brito, I.L. et al. (2019) 'Transmission of human-associated microbiota along family and social networks', Nature Microbiology, 4(6), pp. 964–971. Available at: https://doi.org/10.1038/s41564-019-0409-6.

Browne, H.P. et al. (2021) 'Host adaptation in gut Firmicutes is associated with sporulation loss and altered transmission cycle', Genome Biology, 22(1), p. 204. Available at: https://doi.org/10.1186/s13059-021-02428-6.

Carter, M.M. et al. (2023) 'Ultra-deep sequencing of Hadza hunter-gatherers recovers vanishing gut microbes', Cell, 186(14), pp. 3111-3124.e13. Available at: https://doi.org/10.1016/j.cell.2023.05.046.

Chklovski, A. et al. (2023) 'CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning', Nature Methods, 20(8), pp. 1203–1212. Available at: https://doi.org/10.1038/s41592-023-01940-w.

Chonnacháin, C.N. et al. (2024) 'The effects of dairy on the gut microbiome and symptoms in gastrointestinal disease cohorts: a systematic review', Gut Microbiome, 5, p. e5. Available at: https://doi.org/10.1017/gmb.2024.2.

Danecek, P. et al. (2021) 'Twelve years of SAMtools and BCFtools', GigaScience, 10(2), p. giab008. Available at: https://doi.org/10.1093/gigascience/giab008.

David, L.A. et al. (2014) 'Diet rapidly and reproducibly alters the human gut microbiome', Nature, 505(7484), pp. 559–563. Available at: https://doi.org/10.1038/nature12820.

Debray, R., Tung, J. and Archie, E.A. (2024) 'Ecology and Evolution of the Social Microbiome'. Available at: https://doi.org/10.1146/annurev-ecolsys-102622-030749.

Delgado, L.F. and Andersson, A.F. (2022) 'Evaluating metagenomic assembly approaches for biome-specific gene catalogues', Microbiome, 10(1), p. 72. Available at: https://doi.org/10.1186/s40168-022-01259-2.

Derrien, M. et al. (2022) 'Insights into endogenous Bifidobacterium species in the human gut microbiota during adulthood', Trends in Microbiology, 30(10), pp. 940–947. Available at: https://doi.org/10.1016/j.tim.2022.04.004.

Febinia, C.A. et al. (2022) 'Distinctive Microbiome Type Distribution in a Young Adult Balinese Cohort May Reflect Environmental Changes Associated with Modernization', Microbial Ecology, 83(3), pp. 798–810. Available at: https://doi.org/10.1007/s00248-021-01786-9.

Galperin, M.Y. (2013) 'Genome Diversity of Spore-Forming Firmicutes', Microbiology Spectrum, 1(2), p. 10.1128/microbiolspectrum.tbs-0015–2012. Available at: https://doi.org/10.1128/microbiolspectrum.tbs-0015-2012.

Gomez, A. et al. (2016) 'Gut Microbiome of Coexisting BaAka Pygmies and Bantu Reflects Gradients of Traditional Subsistence Patterns', Cell Reports, 14(9), pp. 2142–2153. Available at: https://doi.org/10.1016/j.celrep.2016.02.013.

Gorvitovskaia, A., Holmes, S.P. and Huse, S.M. (2016) 'Interpreting Prevotella and Bacteroides as biomarkers of diet and lifestyle', Microbiome, 4, p. 15. Available at: https://doi.org/10.1186/s40168-016-0160-7.

Gounot, J.-S. et al. (2022) 'Genome-centric analysis of short and long read metagenomes reveals uncharacterized microbiome diversity in Southeast Asians', Nature Communications, 13(1), p. 6044. Available at: https://doi.org/10.1038/s41467-022-33782-z.

Gupta, V.K., Paul, S. and Dutta, C. (2017) 'Geography, Ethnicity or Subsistence-Specific Variations in Human Microbiome Composition and Diversity', Frontiers in Microbiology, 8, p. 1162. Available at: https://doi.org/10.3389/fmicb.2017.01162.
Jain, C. et al. (2018) 'High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries', Nature Communications, 9(1), p. 5114. Available at: https://doi.org/10.1038/s41467-018-07641-9.
Jari Oksanen et al. (2022) 'vegan: Community Ecology Package.' Available at: https://CRAN.R-project.org/package=vegan.
Johnson, A.J. et al. (2020) 'A Guide to Diet-Microbiome Study Design', Frontiers in Nutrition, 7. Available at: https://doi.org/10.3389/fnut.2020.00079.
Jung, Y. and Han, D. (2022) 'BWA-MEME: BWA-MEM emulated with a machine learning approach', Bioinformatics, 38(9), pp. 2404–2413. Available at: https://doi.org/10.1093/bioinformatics/btac137.
Korpela, K. et al. (2016) 'Intestinal microbiome is related to lifetime antibiotic use in Finnish pre-school children', Nature Communications, 7(1), p. 10410. Available at: https://doi.org/10.1038/ncomms10410.
Krueger, F. (2016) 'Babraham Bioinformatics - Trim Galore'. Available at: https://github.com/FelixKrueger/TrimGalore (Accessed: 20 November 2024).
Kyle Bittinger (2020) 'abdiv: Alpha and beta diversity measures for community ecology.' Available at: https://cran.r-project.org/web/packages/abdiv/index.html.
Leviatan, S. et al. (2022) 'An expanded reference map of the human gut microbiome reveals hundreds of previously unknown species', Nature Communications, 13(1), p. 3863. Available at: https://doi.org/10.1038/s41467-022-31502-1.
Li, H. (2013) 'Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM'. arXiv. Available at: https://doi.org/10.48550/arXiv.1303.3997.
Lu, J. et al. (2023) 'Population-level variation in gut bifidobacterial composition and association with geography, age, ethnicity, and staple food', npj Biofilms and Microbiomes, 9(1), pp. 1–12. Available at: https://doi.org/10.1038/s41522-023-00467-4.
Mallick, H. et al. (2021) 'Multivariable association discovery in population-scale meta-omics studies', PLOS Computational Biology, 17(11), p. e1009442. Available at: https://doi.org/10.1371/journal.pcbi.1009442.
Mancabelli, L. et al. (2017) 'Meta-analysis of the human gut microbiome from urbanized and pre-agricultural populations', Environmental Microbiology, 19(4), pp. 1379–1390. Available at: https://doi.org/10.1111/1462-2920.13692.
Murtagh, F. and Legendre, P. (2014) 'Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?', Journal of Classification, 31(3), pp. 274–295. Available at: https://doi.org/10.1007/s00357-014-9161-z.
Na'im, A. and Syaputra, H. (2012) Kewarganegaraan Suku Bangsa Agama dan Bahasa Sehari-hari Penduduk Indonesia [Nationality, Ethnicity, Religion, and Daily Language of Indonesian Population]. Indonesia: Badan Pusat Statistik [BPS-Statistics Indonesia] (Sensus Penduduk 2010). Available at: https://www.bps.go.id/en/publication/2012/05/23/55eca38b7fe0830834605b35/kewarganegaraan-suku-bangsa-agama-dan-bahasa-sehari-hari-penduduk-indonesia.html (Accessed: 7 November 2024).
Nagata, N. et al. (2022) 'Population-level Metagenomics Uncovers Distinct Effects of Multiple Medications on the Human Gut Microbiome', Gastroenterology, 163(4), pp. 1038–1052. Available at: https://doi.org/10.1053/j.gastro.2022.06.070.
Nakayama, J. et al. (2015) 'Diversity in gut bacterial community of school-age children in Asia', Scientific Reports, 5(1), p. 8397. Available at: https://doi.org/10.1038/srep08397.
Nguyen, L.-T. et al. (2015) 'IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies', Molecular Biology and Evolution, 32(1), pp. 268–274. Available at: https://doi.org/10.1093/molbev/msu300.
Obregon-Tito, A.J. et al. (2015) 'Subsistence strategies in traditional societies distinguish gut microbiomes', Nature Communications, 6(1), p. 6505. Available at: https://doi.org/10.1038/ncomms7505.
Olm, M.R. et al. (2017) 'dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication', The ISME Journal, 11(12), pp. 2864–2868. Available at: https://doi.org/10.1038/ismej.2017.126.
Olm, M.R. et al. (2021) 'inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains', Nature Biotechnology, 39(6), pp. 727–736. Available at: https://doi.org/10.1038/s41587-020-00797-0.
Ondov, B.D. et al. (2016) 'Mash: fast genome and metagenome distance estimation using MinHash', Genome Biology, 17(1), p. 132. Available at: https://doi.org/10.1186/s13059-016-0997-x.
Orakov, A. et al. (2021) 'GUNC: detection of chimerism and contamination in prokaryotic genomes', Genome Biology, 22(1), p. 178. Available at: https://doi.org/10.1186/s13059-021-02393-0.

Pasolli, E. et al. (2019) ‘Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle’, Cell, 176(3), pp. 649-662.e20. Available at: https://doi.org/10.1016/j.cell.2019.01.001.
Penders, J. et al. (2006) ‘Factors Influencing the Composition of the Intestinal Microbiota in Early Infancy’, Pediatrics, 118(2), pp. 511–521. Available at: https://doi.org/10.1542/peds.2005-2824.
Popkin, B.M. (2003) ‘The Nutrition Transition in the Developing World’, Development Policy Review, 21(5–6), pp. 581–597. Available at: https://doi.org/10.1111/j.1467-8659.2003.00225.x.
Ramos-Silva, P., Serrano, M. and Henriques, A.O. (2019) ‘From Root to Tips: Sporulation Evolution and Specialization in Bacillus subtilis and the Intestinal Pathogen Clostridioides difficile’, Molecular Biology and Evolution, 36(12), pp. 2714–2736. Available at: https://doi.org/10.1093/molbev/msz175.
Reyes-García, V. et al. (2019) ‘Dietary transitions among three contemporary hunter-gatherers across the tropics’, Food Security, 11(1), pp. 109–122. Available at: https://doi.org/10.1007/s12571-018-0882-4.
Rosas-Plaza, S. et al. (2022) ‘Human Gut Microbiome Across Different Lifestyles: From Hunter-Gatherers to Urban Populations’, Frontiers in Microbiology, 13. Available at: https://doi.org/10.3389/fmicb.2022.843170.
Rothschild, D. et al. (2018) ‘Environment dominates over host genetics in shaping human gut microbiota’, Nature, 555(7695), pp. 210–215. Available at: https://doi.org/10.1038/nature25973.
Sarkar, A. et al. (2020) ‘Microbial transmission in animal social networks and the social microbiome’, Nature Ecology & Evolution, 4(8), pp. 1020–1035. Available at: https://doi.org/10.1038/s41559-020-1220-8.
Sarkar, A. et al. (2024) ‘Microbial transmission in the social microbiome and host health and disease’, Cell, 187(1), pp. 17–43. Available at: https://doi.org/10.1016/j.cell.2023.12.014.
Schnorr, S.L. et al. (2014) ‘Gut microbiome of the Hadza hunter-gatherers’, Nature Communications, 5(1), p. 3654. Available at: https://doi.org/10.1038/ncomms4654.
Seemann, T. (2014) ‘Prokka: rapid prokaryotic genome annotation’, Bioinformatics, 30(14), pp. 2068–2069. Available at: https://doi.org/10.1093/bioinformatics/btu153.
Sonnenburg, E.D. and Sonnenburg, J.L. (2019) ‘The ancestral and industrialized gut microbiota and implications for human health’, Nature Reviews Microbiology, 17(6), pp. 383–390. Available at: https://doi.org/10.1038/s41579-019-0191-8.
Stražar, M. et al. (2021) ‘Gut microbiome-mediated metabolism effects on immunity in rural and urban African populations’, Nature Communications, 12(1), p. 4845. Available at: https://doi.org/10.1038/s41467-021-25213-2.
Suzuki, T.A. et al. (2022) ‘Codiversification of gut microbiota with humans’, Science, 377(6612), pp. 1328–1332. Available at: https://doi.org/10.1126/science.abm7759.
Swick, M.C., Koehler, T.M. and Driks, A. (2016) ‘Surviving Between Hosts: Sporulation and Transmission’, in Virulence Mechanisms of Bacterial Pathogens. John Wiley & Sons, Ltd, pp. 567–591. Available at: https://doi.org/10.1128/9781555819286.ch20.
Uritskiy, G.V., DiRuggiero, J. and Taylor, J. (2018) ‘MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis’, Microbiome, 6(1), p. 158. Available at: https://doi.org/10.1186/s40168-018-0541-1.
Valdes, A.M. et al. (2018) ‘Role of the gut microbiota in nutrition and health’, BMJ, 361, p. k2179. Available at: https://doi.org/10.1136/bmj.k2179.
Valles-Colomer, M. et al. (2023) ‘The person-to-person transmission landscape of the gut and oral microbiomes’, Nature, 614(7946), pp. 125–135. Available at: https://doi.org/10.1038/s41586-022-05620-1.
Vangay, P. et al. (2018) ‘US Immigration Westernizes the Human Gut Microbiome’, Cell, 175(4), pp. 962-972.e10. Available at: https://doi.org/10.1016/j.cell.2018.10.029.
von Meijenfeldt, F.A.B., Hogeweg, P. and Dutilh, B.E. (2023) ‘A social niche breadth score reveals niche range strategies of generalists and specialists’, Nature Ecology & Evolution, 7(5), pp. 768–781. Available at: https://doi.org/10.1038/s41559-023-02027-7.
Wilson, A.S. et al. (2020) ‘Diet and the Human Gut Microbiome: An International Review’, Digestive Diseases and Sciences, 65(3), pp. 723–740. Available at: https://doi.org/10.1007/s10620-020-06112-w.
Yatsunenko, T. et al. (2012) ‘Human gut microbiome viewed across age and geography’, Nature, 486(7402), pp. 222–227. Available at: https://doi.org/10.1038/nature11053.
Zhernakova, A. et al. (2016) ‘Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity’, Science, 352(6285), pp. 565–569. Available at: https://doi.org/10.1126/science.aad3369.