



RESEARCH PAPER



Unique Pakistani gut microbiota highlights population-specific microbiota signatures of type 2 diabetes mellitus

Afshan Saleem^{a,b,c}, Aamer Ikram^d, Evgenia Dikareva^a, Emilia Lahtinen^a, Dollwin Matharu^a, Anne-Maria Pajari^e, Willem M. de Vos^{a,f}, Fariha Hasan^b, Anne Salonen^g ^{a*}, and Ching Jian ^{a*}

^aHuman Microbiome Research Program, Faculty of Medicine, University of Helsinki, Helsinki, Finland; ^bDepartment of Microbiology, Faculty of Biological Sciences, Quaid-i-Azam University, Islamabad, Pakistan; ^cDepartment of Microbiology, Faculty of Basic and Applied Sciences, University of Haripur, Haripur, Pakistan; ^dDepartment of Virology, National Institute of Health, Islamabad, Pakistan; ^eDepartment of Food and Nutrition, University of Helsinki, Helsinki, Finland; ^fLaboratory of Microbiology, Wageningen University, Wageningen, The Netherlands

ABSTRACT

Biogeographic variations in the gut microbiota are pivotal to understanding the global pattern of host-microbiota interactions in prevalent lifestyle-related diseases. Pakistani adults, having an exceptionally high prevalence of type 2 diabetes mellitus (T2D), are one of the most understudied populations in microbiota research to date. The aim of the present study is to examine the gut microbiota across individuals from Pakistan and other populations of non-industrialized and industrialized lifestyles with a focus on T2D. The fecal samples from 94 urban-dwelling Pakistani adults with and without T2D were profiled by bacterial 16S ribosomal RNA gene and fungal internal transcribed spacer (ITS) region amplicon sequencing and eubacterial qPCR, and plasma samples quantified for circulating levels of lipopolysaccharide-binding protein (LBP) and the activation ability of Toll-like receptor (TLR)-signaling. Publicly available datasets generated with comparable molecular methods were retrieved for comparative analysis of the bacterial microbiota. Overall, urbanized Pakistanis' gut microbiota was similar to that of transitional or non-industrialized populations, depleted in *Akkermansiaceae* and enriched in *Prevotellaceae* (dominated by the non-Westernized clades of *Prevotella copri*). The relatively high proportion of *Atopobiaceae* appeared to be a unique characteristic of the Pakistani gut microbiota. The Pakistanis with T2D had elevated levels of LBP and TLR-signaling in circulation as well as gut microbial signatures atypical of other populations, e.g., increased relative abundance of *Libanicoccus/Parolsenella*, limiting the inter-population extrapolation of gut microbiota-based classifiers for T2D. Taken together, our findings call for a more global representation of understudied populations to extend the applicability of microbiota-based diagnostics and therapeutics.

ARTICLE HISTORY

Received 23 June 2022
Revised 22 September 2022
Accepted 19 October 2022

KEYWORDS

Lifestyle; urbanization; gut microbiota; T2D; metabolic diseases; Toll-like receptor; *prevotella*; random forest


Introduction

The human gut microbiota (i.e., a collection of microbes living in the gut) has been associated with various intestinal and extraintestinal diseases due to its considerable contribution to immune and metabolic homeostasis.¹ Substantial biogeographic variations have been documented in the human gut microbiota, reflecting differences in lifestyle practices, hygiene and pathogen load, diet, medication, and host genetics that altogether influence the assembly and composition of the gut microbiota.² Studies have begun to unravel the impact of these biogeographic variations on health, potentially linking the evolution of the gut microbiota to the varying burden of non-communicable diseases and

the prevalence of infectious diseases in different populations.² Currently, more than two-thirds of public human microbiota datasets originate from Europe and North America,³ whereas a large number of uncultured bacterial species that may have ramifications in health and disease exist in the understudied populations.^{4,5} On the other hand, population studies indicate that the gut microbiota varies between individuals of similar^{6,7} or different ethnicities⁸⁻¹⁰ residing in geographical proximity. Altogether, these intra- and inter-population differences limit the applications of microbiota-based diagnostics and treatments,⁷ which can be exemplified by the debate on microbiota signatures of type 2 diabetes mellitus (T2D). The pathophysiology of

CONTACT Ching Jian  ching.jian@helsinki.fi  Haartmaninkatu 3, PO box 21, FI-00014 Helsinki, Finland

*Equal contribution

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/19490976.2022.2142009>

© 2022 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

T2D, characterized by dysregulated glucose metabolism and insulin resistance, has been suggested to have a microbiota component via low-grade inflammation initiated by pathogen-associated molecular patterns (PAMPs; e.g., lipopolysaccharide (LPS)) and altered levels of short-chain fatty acids (SCFAs; primary saccharolytically derived microbial fermentation products);¹ both mechanisms were predominately inferred from preclinical animal models and associative studies from a few extensively studied populations. However, recent studies have found little convergence of the microbiota characteristics of T2D across geographical regions¹¹ or across cohorts of colocalized individuals of the same ethnicity.¹²

Pakistan as the world's fifth-most populous country had the highest age-adjusted comparative diabetes prevalence in 2021;¹³ Pakistan is experiencing rapid urbanization with increasing consumption of the Western diet and lifestyle changes, especially in the urban areas,¹⁴ which all have implications for the gut microbiota. However, virtually nothing is known about the gut microbiota of Pakistanis in a global context, as the population of 220 million is represented by <100 samples from a few pilot studies.^{14,15} In general, Southern and Central Asia is the most underrepresented region for microbiota research.³ To bridge the knowledge gap, we aimed to profile the gut microbiota of Pakistani adults with and without T2D and to assess the generalizability of the gut bacterial signature of T2D across different cohorts. We systematically compared the healthy Pakistani microbiota to that of industrialized (represented by China, Japan, and Finland) and transitional or non-industrialized (represented by Indonesia, Sudan, and rural India) populations as well as that of Pakistani adults with T2D. Potential differences in the butyrate production capacity of the gut microbiota, circulating levels of lipopolysaccharide-binding protein (LBP), and Toll-like receptor (TLR)-signaling in circulation between Pakistani participants with and without T2D were analyzed to gain mechanistic insights. We then evaluated whether a classification model of T2D based on the Pakistani gut microbiota can be extrapolated to other populations. Only publicly available datasets targeting the V3-V4 region of the 16S rRNA gene and employing the Illumina sequencing

technology were included in the present study to minimize methodological artifacts, as previous studies suggest that using matching variable regions of the 16S rRNA gene represents one of the most essential factors for cross-study comparability.^{16,17}

Results

Taxonomic characteristics of the Pakistani gut microbiota in a global context

Our Pakistani cohort included 94 urban-dwelling adults living in and around the capital regions with and without confirmed T2D. Their gut microbiota was profiled using 16S rRNA gene amplicon sequencing targeting the V3-V4 region, which resulted in $17,160 \pm 910$ quality-controlled and chimera-checked reads per sample. Overall, Firmicutes and Actinobacteria were the most abundant phyla in the Pakistani population accounting for 63.7% and 25.2% of the total read counts on average, respectively, and they were observed in all the samples. Bacteroidetes and Proteobacteria constituted 5.3% and 4% of the total read counts, respectively, yet these microbes were found in 87% and 88% of the study participants, respectively. Verrucomicrobia (genus *Akkermansia*) made up less than 1% of the gut bacterial community and was detectable in only 16% of the individuals. Spirochetes (dominated by *Spirochaetaceae/Treponema* 2), a phylum observed mainly in ancient and non-industrialized societies but absent in industrialized populations,¹⁸ were detected in 10 Pakistani participants with an average abundance of 0.13%.

Besides a pilot study showing a few phylum-level differences when comparing urban Pakistani adults' gut microbiota to the now defunct uBiome database of unspecified origin,¹⁴ the urban Pakistanis' gut microbiota has not been contextualized on a global scale. Therefore, we compared the healthy participants' microbiota to their counterparts from publicly available datasets generated with comparable molecular methods for profiling the fecal microbiota, including an external cohort of Pakistanis sampled from the same area (Table S1). Visualization of the Bray-Curtis dissimilarity measure using Principal Coordinates Analysis (PCoA) revealed a separation of industrialized populations (China, Japan, and Finland) from the rest along the primary axis ($P = .001$, PERMANOVA; Figure 1a). Moreover, the

East Asian populations clustered tightly, slightly separated from the Finnish cohort, mirroring a recent study showing the impact of East Asian ethnicity on the gut microbiota.¹⁰ The overall structure of the Pakistani gut microbiota in our cohort (Pakistan1/

PAK1) overlapped with that of the Indonesian gut microbiota but somewhat differed from the other Pakistani cohort (Pakistan2/PAK2). The rural Indian population occupied a separate corner in the ordination space. The two Pakistani cohorts were similar in

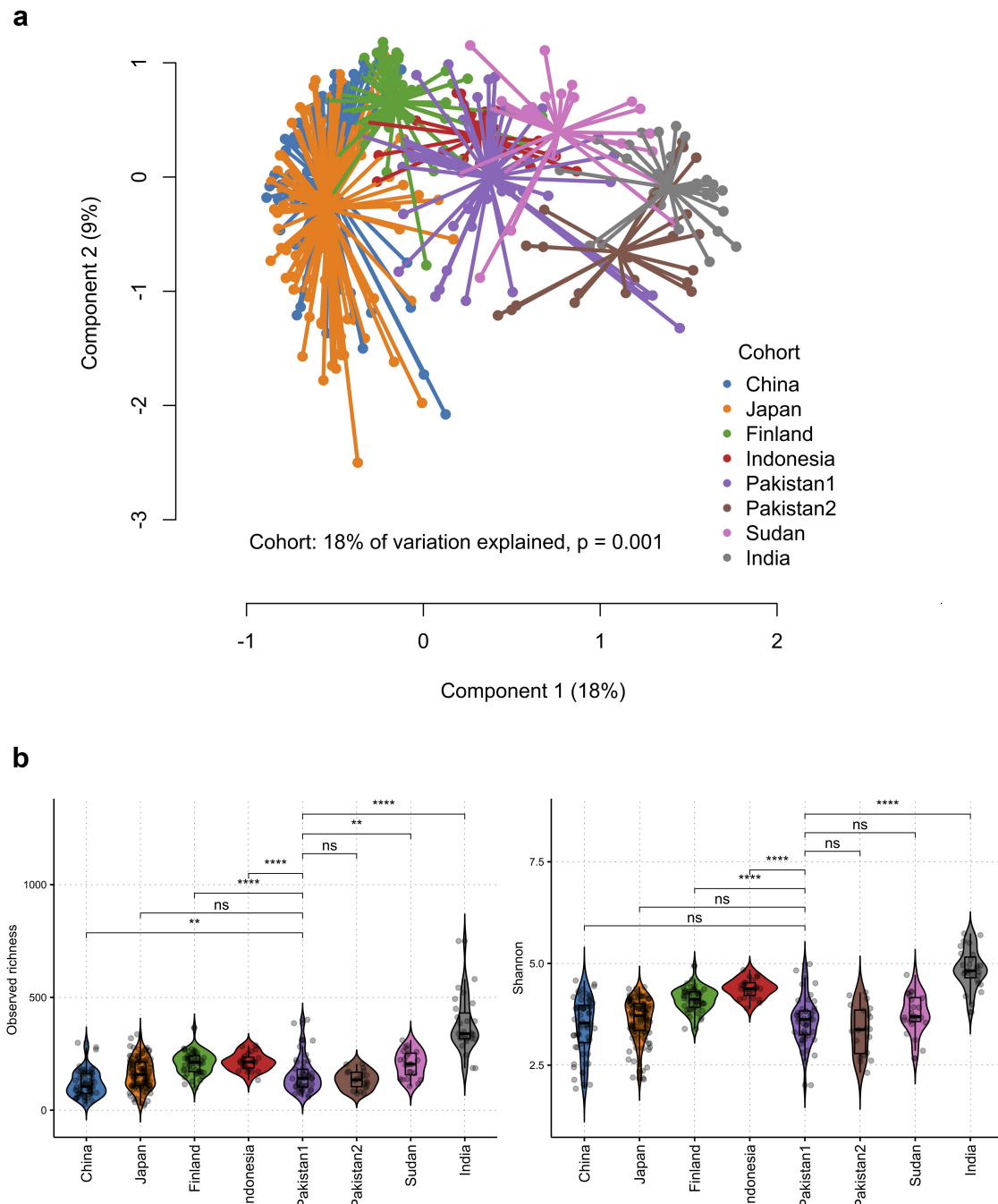


Figure 1. (a) Principal coordinate analysis (PCoA) plot of microbiota variation from different cohorts of healthy adults based on the Bray-Curtis dissimilarity matrix. Cohort explained 18% of the microbiota variation ($P = .001$, PERMANOVA). (b) Violin plots (a combination of the box plot with a kernel density plot) showing microbiota α -diversity (observed richness and Shannon's diversity index) per each cohort. The center line denotes the median, the boxes cover the 25th and 75th percentiles, and the whiskers extend to the most extreme data point, which is no more than 1.5 times the length of the box away from the box. Points outside the whiskers represent outlier samples. Significance was calculated by the Wilcoxon rank-sum test using Pakistan1 as the reference. **** $P < .0001$; *** $P < .001$; ** $P < .01$; * $P < .05$; "ns" $P > .05$.

microbiota α -diversity metrics and generally were on the level of that of the Chinese and Japanese cohorts that had the lowest α -diversity (Figure 1b). The Indian cohort had the highest microbiota α -diversity as previously reported for other non-industrialized populations.² In terms of taxonomic composition of the gut microbiota on the family level, stark differences between the industrialized and transitional/non-industrialized populations existed in *Akkermansiaceae*, *Enterobacteriaceae*, and *Prevotellaceae*; *Prevotellaceae* and *Enterobacteriaceae* were dominant in both Pakistani cohorts and *Akkermansiaceae* represented as an absentee, similar to other non-industrialized populations (Figure 2; Fig. S1). The high prevalence and abundance of *Atopobiaceae* (dominated by *Olsenella* and *Libanicoccus*) was uniquely observed in the Pakistani gut microbiota (Fig. S1). Other taxonomic features commonly found in both Pakistani cohorts included highly abundant *Bifidobacteriaceae* (dominated by *Bifidobacterium*) and *Erysipelotrichaceae* (dominated by *Catenibacterium* and *Holdemanella*) as well as an exceptionally high proportion of *Streptococcaceae* in some individuals (Figure 2; Fig. S1). *Spirochaetaceae*

(belonging to Spirochetes) could be detected only in the Pakistanis (11% prevalence in PAK1 and 15% in PAK2), Sudanese (31% prevalence), and Indians (7% prevalence).

Prevotella copri of the *Prevotellaceae* family can be grouped into four clades (A, B, C, and D) depending on their genetic structure; clade A was ubiquitously found among both the Westernized and non-Westernized populations, while clades B, C, and D were predominantly found in non-Westernized populations.¹⁹ As the urban Pakistanis' gut microbiota appears to exhibit features of the non-industrialized gut microbiota, we evaluated whether the urban Pakistani population has *P. copri* clades typical of non-Westernized populations. Consequently, we found that *P. copri* ASVs in both Pakistani cohorts belonged to clades B and C (Table S2).

Taxonomic, functional and molecular signatures in the gut microbiota of Pakistanis with T2D

Focusing on our PAK1 cohort, age and sex were similarly distributed for the T2D patients and



Figure 2. Heat map showing the relative abundance of top 20 dominant bacterial families per cohort.

healthy controls and for the patients prescribed with metformin (metformin users) and insulin therapy (non-metformin users). BMI, total cholesterol (TC), triglycerides (TG), and HbA1c were significantly higher in those with T2D or the non-metformin users ($P < .05$, Table 1). The individuals with and without T2D had significant differences in self-reported diets. Among the variables mentioned above, triglycerides and HbA1c significantly explained the variation in the gut microbiota in

the model as determined by *EnvFit* ($P < .01$, Table 1).

Principal coordinate analysis (PCoA) showed a clear separation between the gut microbiota of patients and controls (Figure 3), demonstrating that T2D status was associated with the gut microbiota composition, explaining 9% of the total variation ($P < .001$, PERMANOVA). There was no difference in the observed richness, but Shannon's diversity and eubacterial density were drastically

Table 1. Anthropometric and biochemical parameters and dietary characteristics of participants and their associations with the gut microbiota based on β -diversity.

	Control (N = 46)	T2D (N = 48)	Association with β -diversity			Met- (N = 21)	Met+ (N = 27)	P
	Mean/ median	Mean/ median	P	R ²	P	Mean/ median	Mean/ median	
Basic characteristics								
Age	49.0	51.1	>0.05	0.008	>0.05	54.0	48.9	>0.05
Sex			>0.05	0.012	>0.05			>0.05
BMI	25.0	27.5	<0.01	0.047	>0.05	30.2	25.5	<0.01
TC	170.0	177.5	<0.01	0.018	>0.05	190.0	167.0	<0.05
TG	132.5	245.0	<0.01	0.11	<0.01	370.0	185.0	<0.01
HbA1c	5.3	7.5	<0.01	0.3	<0.01	7.8	7.3	<0.01
Diet (frequency)								
Vegetables and fruits	19.5	28.4	<0.01	0.08	<0.05	28.8	28.1	>0.05
Meat	11.6	6.2	<0.01	0.08	<0.05	6.0	6.4	>0.05
Fast food and deep fried food	2.8	0.7	<0.01	0.04	>0.05	0.2	1.1	<0.05

BMI, body mass index; TC, total cholesterol; TG, triglycerides; HbA1c, hemoglobin A1c; T2D, type 2 diabetes mellitus; Met-, T2D patients prescribed with insulin therapy; Met+, T2D patients prescribed with metformin.

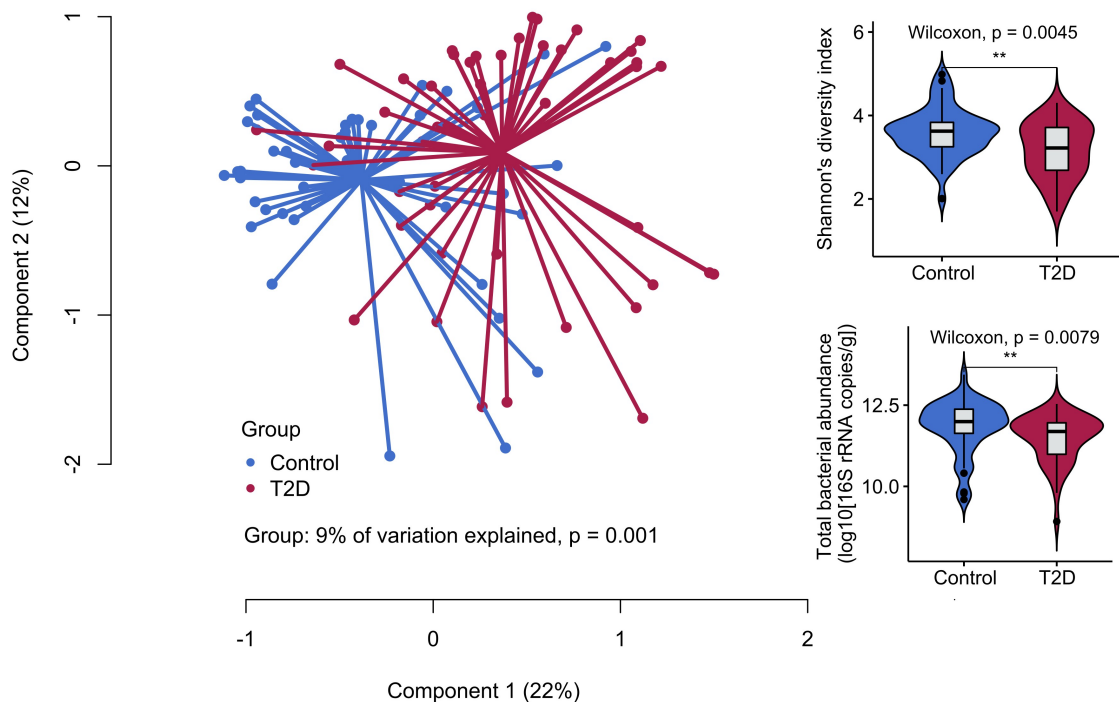


Figure 3. Significant differences in microbiota β -diversity assessed by Principal coordinates analysis (PCoA) plot and gut microbiota ecology represented by Shannon's diversity index and eubacterial density between Pakistani adults without (blue) and with (red) T2D. **** $P < .0001$; *** $P < .001$; ** $P < .01$; * $P < .05$; "ns" $P > .05$.

reduced in T2D patients, suggesting a strongly aberrant community ($P < .001$, **Figure 3**).

To identify taxa and functional modules driving the microbiota differences, we compared the abundance of individual bacterial families, genera, species, and imputed KEGG pathways between the patients and controls using a negative binomial generalized linear model, as implemented in differential expression analysis for sequence count data version 2 (*DESeq2*). The analyses were additionally controlled for BMI and diet to disentangle their potential confounding effects. Striking differences (all FDR-adjusted $P < .05$) between the groups were found on all taxonomic ranks as well as in the functional modules with and without controlling for covariates. On the family level (**Figure 4a**), *Lactobacillaceae* and *Coriobacteriaceae* were

enriched and *Ruminococcaceae* depleted in T2D patients, which remained significant after adjusting for BMI and diet. Moreover, a marked decrease in *Prevotellaceae* (mainly attributable to *Prevotella 9/Prevotella copri*) was noted, which lost significance after controlling for BMI. On the genus level (**Figure 4b**), T2D patients had consistently increased proportions of *Libanicoccus*, *Lactobacillus*, *Collinsella*, *Senegalimassilia*, *Bifidobacterium*, and *Slackia* and reduced relative abundances of *Faecalibacterium* and *Oribacterium*. The relative abundance of *Collinsella* has been shown to correlate with elevated circulating insulin,²⁰ increased gut permeability,²¹ and altered bile acid metabolism²² that contribute to the pathophysiology of T2D. *Lactobacillus* represents one of the most discrepant signatures of the T2D gut

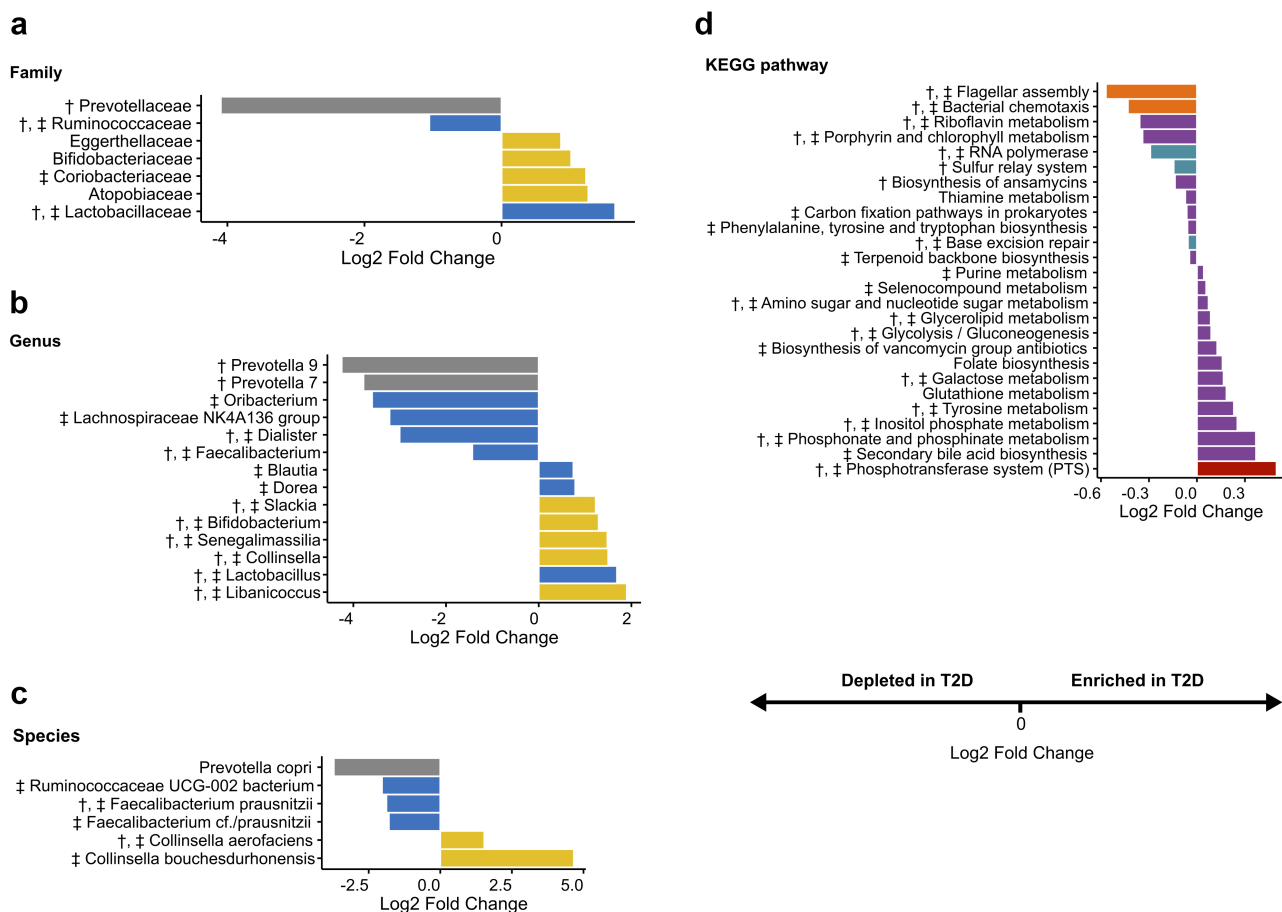


Figure 4. Differentially abundant bacterial (a) families (b) genera (c) species (d) KEGG pathways (level 3) visualized by divergent bar plots. Only statistically significant results are shown (FDR-adjusted $P < .05$). Log₂ fold change was calculated using controls as the reference group. Bacterial taxa are colored according to their respective phyla (Bacteroidetes in gray, Actinobacteria in blue, and Firmicutes in yellow). KEGG pathways are colored at level 1 (Environmental Information Processing in red, Metabolism in purple, Genetic Information Processing in teal, and Cellular Processes in Orange). The differential bacterial taxa and KEGG pathways that remained significant after controlling for BMI or diet type are marked with † and ‡, respectively.

microbiota, likely due to its high genomic diversity.²³ Species were assigned to the ASVs that perfectly matched reference sequences, among which *Collinsella bouchesdurhonensis* and *Collinsella aerofaciens* were consistently enriched and *Faecalibacterium prausnitzii* depleted in T2D patients (Figure 4c). The alternations in the bacterial taxa were expected to result in extensive changes in the metabolic potential of the gut microbiota, with 20 out of 26 KEGG pathways categorized as “metabolism” (Figure 4d). Notably, the modules related to carbohydrate metabolism (i.e., inositol phosphate metabolism, galactose metabolism, glycolysis/gluconeogenesis, and amino sugar and nucleotide sugar metabolism), amino acid-related metabolism (i.e., phosphonate and phosphinate metabolism, tyrosine metabolism, glutathione metabolism, and selenocompound metabolism), and lipid metabolism (i.e., secondary bile acid biosynthesis and glycerolipid metabolism) were significantly more abundant in T2D patients (Figure 4d). The bacterial phosphotransferase system (PTS), which mediates the uptake of multiple sugars from the environment,²⁴ was predicted to be higher in its capacity appreciably in the T2D gut microbiota (Figure 4d). This, and perturbations in the functional modules related to amino acid metabolism, concurred with the findings from the previous studies in Indian-Danish prediabetic and T2D patients.^{11,25}

We additionally profiled the fungal communities using ITS1 sequencing, generating 7560 ± 785 quality-filtered ITS1 sequences per sample. Nevertheless, the majority of the pre-processed sequences failed to be taxonomically annotated,

leaving only 28 participants with gut fungal profiles (control = 11, T2D = 17). The large fraction of unannotated reads suggests the existence of novel species from the Pakistanis not yet in sequence repositories. No difference in fungal β -diversity was found between the 28 participants with and without T2D (Fig. S2A). While the taxonomic composition was highly variable (Fig. S2B), *Candida sake*, *Teunomyces*, and *Candida akabanensis* were identified as significantly enriched in the patients with T2D (all FDR-adjusted $P < .05$, Fig. S2C).

We next assessed the molecular links between the gut microbiota and T2D, specifically butyrate as a major SCFA and PAMPs that have been suggested to be the cardinal microbial mediators in metabolic disorders.¹ Butyrate production capacity estimated by quantifying the butyryl-CoA:acetate CoA-transferase gene (responsible for a major route for butyrate production in bacteria) was unexpectedly higher in T2D patients ($P < .001$, Figure 5a). An increased influx of PAMPs into circulation, especially LPS (a component of the Gram-negative bacterial outer membrane), has been linked to inflammation and impaired glucose metabolism through activation of TLR4 or TLR2-dependent signaling.²⁶ Thus, we measured the potential of plasma samples to activate innate immunity receptors TLR4 (receptor for LPS) and TLR2 (receptor for a variety of Gram-negative and Gram-positive bacterial products). The plasma samples from T2D patients induced significantly higher levels of TLR4 activation ($P < .001$, Figure 5b), and elevated levels of TLR2 activation by trend ($P = .061$, Figure 5c). For LPS-TLR4 signaling, LPS-binding protein (LBP) attaches to LPS and presents it to CD14 to initiate TLR4 activation; we subsequently quantified the

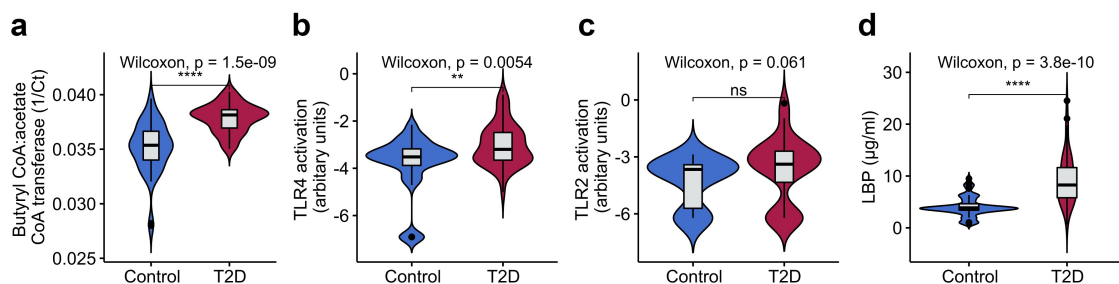


Figure 5. Violin plots showing differences in (a) butyrate production capacity (b) Toll-like receptor (TLR) 4 activation (c) TLR2 activation and (d) lipopolysaccharide binding protein (LBP) between controls and T2D patients. The center line denotes the median, the boxes cover the 25th and 75th percentiles, and the whiskers extend to the most extreme data point, which is no more than 1.5 times the length of the box away from the box. Points outside the whiskers represent outlier samples. Significance was calculated using the Wilcoxon rank-sum test. **** $P < .0001$; *** $P < .001$; ** $P < .01$; * $P < .05$; “ns” $P > .05$.

concentrations of plasma LBP and found increased levels of LBP in T2D patients ($P < .001$, Figure 5d), mirroring the findings on TLR4 activation. Taken together, these findings suggest that Pakistani T2D patients possess an increased potential for butyrate production in the gut microbiota and engage in elevated TLR-signaling in circulation.

Since approximately half of the patients with T2D were prescribed with metformin in our cohort (Table 1), we conducted subgroup analysis for the patients to identify potential effects of metformin on the gut microbiota. Consequently, no statistically significant differences in microbiota α - and β -diversity, eubacterial density, butyrate production capacity, TLR2 and 4 activation, and the relative abundances of individual bacterial genera were identified between the groups (Fig. S3). A significantly lower level of plasma LBP was found in the patients treated with metformin ($P = .02$, Fig. S3B).

Correlations between metabolic parameters, bacterial taxonomic features, and microbial mediators

We assessed all triplets of pairwise interactions between host metabolic phenotypes, bacterial genera, and microbial mediators using Spearman correlations (requiring FDR-adjusted $P < .05$ in each comparison of two data spaces). Figure 6 shows a chord diagram constructed from these data, where the bacterial genera enriched in T2D (e.g., *Libanicoccus*, *Collinsella*, and *Senegalimassilia*) were positively and members of *Prevotellaceae* negatively associated with plasma LBP and/or negative metabolic variables. Two significant correlations were identified between TLR4 activation and bacterial genera: *Faecalibacterium* (reduced in T2D patients) was negatively and *Libanicoccus* (enriched in T2D patients) was positively associated with TLR4 activation (Figure 6). The aforementioned associations remained significant in the subgroup analysis of T2D patients (Fig. S4).

Classification models of T2D based on bacterial taxonomic signatures and their cross-study portability

Varying and sometimes contrasting microbiota signatures of T2D have been documented in different cohorts and populations,^{23,27} and their

cross-study portability as biomarkers has been questioned.¹¹ Given that we observed considerable taxonomic differences in the microbiota between the controls and T2D patients, many of which unreported previously, e.g., *Libanicoccus* and *Slackia*, a random forest (RF) classifier model was constructed using the prevalent bacterial genera (prevalence $>10\%$) detected in our cohort (Table S3) that could specifically identify T2D patients with high overall accuracy (area under curve; AUC = 0.8684, Figure 7a). As a conservative approach, this RF classifier was trained excluding the patients prescribed with metformin to avoid potential effects of metformin on the gut microbiota^{11,28} unobserved in our *DESeq2* analysis. Next, we performed study-to-study model transfer (model trained for each study independently and their predictive performance evaluated on the other datasets) to systematically assess cross-study generalization of T2D microbiota signatures. The RF models for other cohorts of previously published studies were trained without excluding those treated with metformin, as the records of medication were unavailable on the individual level. Despite the conservative approach, population-specific effects were evident where the extrapolation of T2D classifiers performed satisfactorily among the Pakistanis, and to a lesser extent, the Sudanese (median AUC = 0.71, range = 0.62–1.0; Figure 7b); relatively poor discriminatory power was pervasive for other populations (AUC < 0.7 , Figure 7b). To understand the population-specific discrimination accuracy, we applied *DESeq2* across different studies/cohorts to top 10 predictive microbial features extracted from the RF model trained on our cohort (Figure 7c). The directionality of differences in these features between controls and T2D patients was largely congruent among the two Pakistani and the Sudanese cohorts, but inconsistent or discordant at times in other cohorts (Figure 7c). Of note, *Prevotella* 9, a highly discordant signature between the Pakistani and Chinese cohorts, was dominated by different clades of *Prevotella copri*; clade A was predominately present in the Chinese, while clades B and C were exclusively found in the Pakistanis (Table S2). Interestingly, *Libanicoccus* and *Slackia* appear to be the

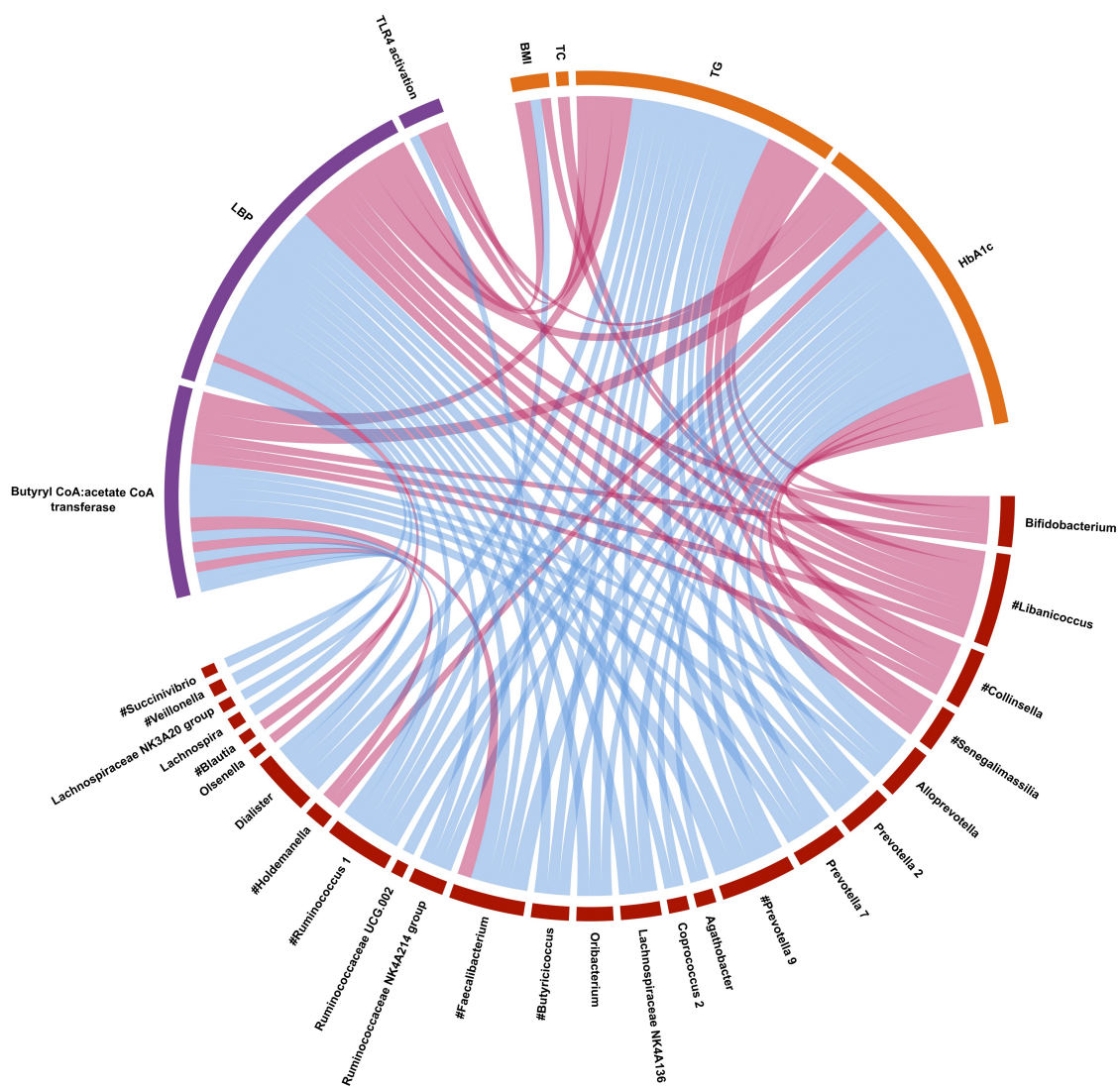


Figure 6. A Chord diagram visualizing the significant interrelation between host metabolic characteristics (cells in Orange) and microbiota taxonomic features (cells in burgundy), and microbial mediators (cells in purple). Features are shown that form triplets of phenotype, microbial and microbial molecule-related variables where at least two of three correlations are significant (Spearman FDR-adjusted $P < .05$). Color of the connectors indicates positive (in red) or negative (in blue) Spearman's rho values. The bacterial genera consistent with the subgroup analysis of T2D patients (Fig. S4A) are marked with #. BMI, body mass index; TC, total cholesterol; TG, triglycerides; HbA1c, hemoglobin A1c; LBP, lipopolysaccharide binding protein; TLR4, Toll-like receptor 4.

taxonomic signatures of T2D specific to Pakistanis, and the two taxa were prevalent in Pakistanis (Fig. S5). *Dialister* as a contrasting signature between the two Pakistani cohorts belongs to the few bistable taxa that differ in bimodal distribution patterns across cohorts, potentially indicating underlying alternative states associated with host factors.²⁹ This cohort-specific bimodality of *Dialister* was observed between the two Pakistani cohorts (Fig. S6), suggesting the presence of cohort-specific characteristics (e.g., medical conditions unaccounted for

in the two studies) in the same population that contributed to the contrasting signature. Of note, some differences in the gut microbiota between the two Pakistani cohorts of healthy controls were apparent (Figures 1 & 2; Fig. S1).

Discussion

The widespread changes in lifestyle and diet due to rapid urbanization in Pakistan over the last two decades have been linked to increased incidence of non-communicable diseases.^{30,31} Urbanization and

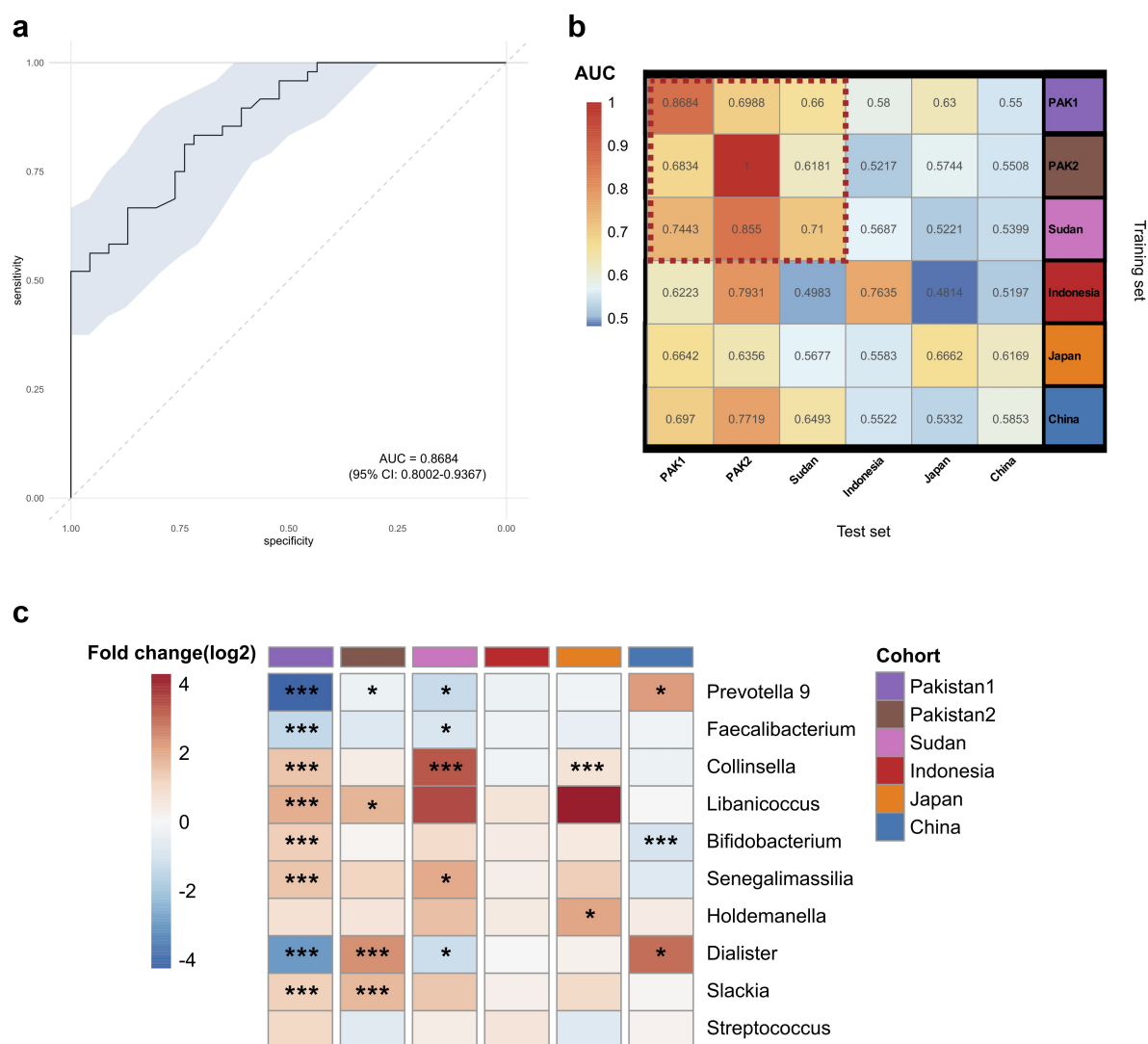


Figure 7. (a) A receiver operating characteristic (ROC) curve evaluating the ability to distinguish T2D cases from controls using a random forest classifier trained on bacterial genus-level abundances, achieving an area under ROC curve (AUC) value of 86.84% with 95% CI of 80.02% to 93.67%. (b) T2D classification accuracy resulting from cross-validation within each study i.e., the test and training set are the same study (the boxed along the diagonal) and study-to-study model transfer (external validations off the diagonal) as measured by the AUC for the classification models trained on genus-level abundances. The color of the scale bar on the right represents the AUC value. The block of cohorts where the classifier had higher discriminatory power is encircled by red dotted lines. (c) *DESeq2* outputs comparing top 10 taxonomic features with the highest discriminatory power extracted from the random forest model trained on the present cohort (PAK1/Pakistan1) between controls and patients with T2D. Log₂ fold change was calculated using controls as the reference group and significance was determined by *DESeq2*. **** P < .0001; *** P < .001; ** P < .01; * P < .05.

industrialization may remodel a population's gut microbiota via multiple exposures, such as Westernization of diet, increased antibiotic use, pollution, improved or deteriorated hygiene status, and early-life microbial exposure.^{32,33} These selective forces could deviate the gut microbiota further from its ancestral state via intergenerational transmission.³⁴ In some studies, the microbiota associated with humans in the industrialized world has been

characterized by progressive disappearance of the *Prevotellaceae*, *Spirochaetaceae*, and *Succinivibrionaceae* families, and increased abundance and prevalence of *Akkermansia* and *Bacteroides*.^{2,18} Our results recapitulate these observations and place the urban Pakistani gut microbiota into the transitional or non-industrialized category. The persistence of non-industrialized microbiota signatures is in agreement with previous reports on transitioning populations or

communities, such as first-generation migrants in the Netherlands,⁸ Irish Travelers enforced to abandon nomadism,³⁵ and urban Nigerians.³⁶ In contrast, a study about Thai migrants moving to the United States claimed that the microbiota rapidly assumes the structure of the new place of residence.³⁷ While the degree and timeframe to which the microbiota signatures persist following lifestyle modernization remains an open question, transitioning from the non-industrialized to industrialized gut microbiota has been suggested to negatively affect metabolic health.^{35,37} It remains unclear what caused the lack of detectable levels of *Akkermansia* in our Pakistani cohort and other non-industrialized populations. Nevertheless, it is possible that the absence of *Akkermansia* in the urban Pakistani adults is maladaptive, as this bacterium provides ecosystem services needed to maintain metabolic homeostasis especially in humans on the Western diet.¹ Further research therefore is warranted to establish the causality between the transitioning gut microbiota and worsening metabolic health in developing countries. While previous research has generally agreed that the gut microbiota of industrialized populations have decreased biodiversity,^{18,38} the gut microbiota of both urban Pakistani cohorts had relatively low biodiversity in the transitional/non-industrialized category of our study. Similarly, the microbiota diversity appeared highly variable (in some cases nearly indistinguishable from that in a US cohort) within non-industrialized African populations and hence cannot be explained solely by different subsistence lifestyles.³⁹ A metagenomic study in rural and recently urbanized Chinese adults found that urbanization was associated with the loss of microbiota diversity and suppression of novel human symbionts,⁴⁰ potentially ascribable to several factors that limit the dispersal of symbionts in urban areas including sanitation, water treatment, and medication. However, our study was not designed to analyze the individual factors that potentially explain the low microbiota diversity in the urban Pakistani participants.

Regarding the population-specific features of the Pakistani gut microbiota, our results are in agreement with a pilot study in a cohort of 32 urban Pakistani adults targeting the V4 region of the 16S rRNA gene.¹⁴ For example, the enrichment of *Bifidobacterium* and *Lactobacillus* is likely attributable to the frequent consumption of fermented

foods in Pakistan;¹⁴ the occasionally high loads of *Enterobacteriaceae* and *Streptococcaceae* that contain several opportunistic pathogens likely result from deteriorated water and hygiene conditions during galloping urbanization. Data from our comparative analysis also reveal the high prevalence and abundance of *Prevotella* 9 in Pakistanis dominated by the non-Westernized clades of *Prevotella copri*, commonly thought to associate with plant-rich diets that are abundant in carbohydrates and fibers. The Pakistani diet is particularly low in intakes of fiber-rich fruits, vegetables, nuts, and whole grains on the global scale.⁴¹ Also, there was no difference in the relative abundance of *Prevotella* between the Pakistani participants who consumed plant-based or meat-based diets (Fig. S7). Thus, the dominance of *Prevotella* in Pakistanis may have been promoted by non-dietary factors. This is supported by the study in Irish Travelers, a historically nomadic ethnic group of European ancestry, who consume a Western-like diet rich in fat and protein with little fiber intake.³⁵ *P. copri* was found to be enriched to the same levels of other non-industrialized populations in a subgroup of Irish Travellers who maintained traditional lifestyles and living conditions, while no difference in diet was found between the subgroups of Irish Travelers.³⁵ A recent study in rural Gambian infants found that *P. copri* rises to dominance rapidly in the first year of birth and remains the most abundant species during the remainder of early childhood,⁴² suggesting its establishment is strongly affected by early-life factors in *Prevotella*-rich populations. In addition to longer and more common exclusively breastfeeding, most weaning infants in Pakistan are initially given unprocessed foods as opposed to convenience baby foods given to infants in the Western world;⁴³ solid foods introduced to most Pakistani infants are predominantly plant-based and family foods,⁴³ particularly those made from roots and tubers that constitute a rich source of resistant starch.⁴⁴ Taken together, future work should address the kinetics of colonization and development of *Prevotella* in relation to various dietary and non-dietary exposures in Pakistani infants.

As expected, both discrepancies and commonalities are noted regarding the taxonomic signatures of T2D between our findings and previous studies in various populations. The depletion of anti-

inflammatory *Faecalibacterium prausnitzii* in T2D has been relatively consistent across studies.^{23,25} Here, we show a negative correlation between the abundance of *Faecalibacterium* in the fecal microbiota and TLR4 activation by the plasma, recently also reported for gut epithelium in an *in vitro* study.⁴⁵ The enrichment of several members of the *Coriobacteriia* class (i.e., *Libanicoccus*, *Senegalimassilia*, and *Slackia*) in Pakistani T2D patients, uncommon in the Western populations, appears to be novel population-specific signatures. *Libanicoccus massiliensis* was isolated from the feces of a healthy Congolese pygmy woman in 2018,⁴⁶ but has since been proposed to be reclassified as a member of the *Parolsenella* genus.⁴⁷ *Slackia* has been sparsely described in the industrialized world, but relatively abundant in traditional societies, e.g., rural Papua New Guineans.³⁸ Little is known about *Libanicoccus/Parolsenella* as a newly isolated bacterium. A recent re-analysis of the studies in experimental autoimmune encephalomyelitis (EAE), the animal model of multiple sclerosis, associated *Parolsenella catena* with the development of pathology in marmosets;⁴⁸ a metagenomic study in rural China reported that *Libanicoccus* was associated with arterial plaque buildup.⁴⁹ The positive relationship between *Libanicoccus/Parolsenella* and TLR4 activation shown in the present study suggests a pro-inflammatory state in T2D potentially mediated by this bacterium.

Importantly, the aforementioned taxa belonging to *Coriobacteriia* are closely related, since the genome-based boundaries between its members are tenuous.^{4,47} Almeida *et al.* reported that a large number of unclassified near-complete metagenomic species (i.e., potential new families and/or genera), most frequently assigned to the *Coriobacteriaceae* family, can be found in the gut microbiota data outside North America and Europe.⁴ Thus, our study highlights the importance of sampling underrepresented populations and regions to uncover novel microbes relevant for human health and improve the applicability of microbiota-based diagnostic strategies. The latter is further underscored by our findings on T2D classification models, where the discriminative power fell short when extrapolating a classification model to other populations. As the discriminative power of machine learning-based classification models is

heavily influenced by the presence/absence of specific microbial taxa,⁵⁰ additional variability from a greater pool of heterogeneous samples from various populations could enable classifiers to capture specific signatures while minimizing overfitting on idiosyncrasies of a single population. The potential of this strategy in improving microbiota-based disease classification has been demonstrated in recent proof-of-concept studies.^{27,51}

The patterns of *Prevotella* and *Bifidobacterium* in prediabetes and T2D are highly inconsistent across cohorts as shown in our and other studies.^{25,52,53} Higher abundance of *P. copri* in the gut microbiota has been suggested to be both conducive and detrimental to glucose homeostasis and host metabolism in different studies, and its role in human health is still under investigation.⁵⁴ Tett *et al.* found no association between the prevalence and abundance of the four *P. copri* clades and the etiology of several diseases, including T2D.¹⁹ However, the two cohorts of T2D included in the Tett *et al.*'s study are from the Chinese and European populations that generally have low prevalence of *P. copri* dominated by clade A. Therefore, the potential clade-specific effects of *P. copri* in T2D require further investigation in *Prevotella*-rich populations with non-Westernized *P. copri* clades using metagenomic sequencing. Additionally, we propose a scenario where some of the observed microbiota signatures in Pakistani T2D patients, namely the reduction in *Prevotella* and expansion in *Bifidobacterium* and butyrate production capacity, are reflective of their transitioning gut microbial ecosystem toward the industrialized one. Supporting this, the microbiota profiling of non-industrialized/non-urbanized African communities suggests a different ecological layout supporting SCFA production with little or no contribution from *Bifidobacterium* and typical butyrate producers,^{36,55} corresponding to their "high-propionate and low-butyrate" SCFA profile.⁵⁵ In general, the effects of *Bifidobacterium* and butyrate in promoting human health have been understudied in non-Western populations, where human-microbe associations may need to be interpreted in an entirely different context.³⁶

A growing body of studies claims to have identified increased microbial components (e.g., LPS) in T2D, potentially contributing to its pathogenesis.⁵⁶

These studies have nonetheless been challenged by various issues, including high technical variability and inability to differentiate stimulatory and inhibitory forms of LPS.^{57,58} By directly quantifying the ability of plasma samples to elicit TLR-signaling, here we corroborated the elevated levels of circulating pro-inflammatory mediators in T2D patients. To our knowledge, this approach has not been applied to human blood samples. It is worth noting that other non-microbially derived molecules can also contribute to the inflammatory response by activating TLR4 signaling,⁵⁹ which may partly explain a few associations between the gut microbiota and TLR4 activation. Importantly, the gut microbial signatures do not necessarily translate to pro-inflammatory potential, which can be conceived as a net effect of pro-inflammatory microbial mediators, translocation rate of the mediators and host's ability to handle them.

Our comparative analysis is limited by sample size owing to missing metadata or restricted data availability in many published studies;⁶⁰ the available metadata for all the 944 samples used in the present study has been included in Table S4. We were nonetheless able to recapitulate critical observations made by previous studies, e.g., the traits of the non-industrialized gut microbiota.¹⁸ Moreover, the unique characteristics of the Pakistani gut microbiota were further validated in an independent cohort. On the other hand, our study focusing on urbanized Pakistanis of a similar social background was unable to fully capture the full diversity of Pakistan. To substantiate our findings and hypothesis on the transitioning gut microbiota, future large studies should expand to peri-urban and rural communities. Although we did not observe the effects of metformin on the gut microbiota previously reported in Chinese, Japanese, and European cohorts of T2D patients,⁶¹ the beneficial effects of metformin have been suggested to be partially mediated by the gut microbes, such as *P. copri*⁶² and *Akkermansia muciniphila*⁶³ that have different abundance and prevalence distributions in Pakistanis. Therefore, further research is warranted to assess the potential contribution of the gut microbiota to the variability in the therapeutic response of metformin in the Pakistani population. Lastly, by selecting and processing the datasets generated by comparable library

preparation protocols and sequencing technology, DNA extraction methods arguably represent the main source of technical variation in our study (Table S1). Nevertheless, the technical variation is minor in comparison to the biological variation in the microbiota when a mechanical disruption step is introduced during DNA extraction^{17,64} and hence that was used as an inclusion criterion to select the datasets for the present study (Table S1).

In conclusion, we show that the urbanized Pakistanis' gut microbiota retains non-industrialized features with several distinctive traits. These unique microbiota characteristics extend to T2D-associated signatures, potentially reflecting transitioning gut microbial ecosystems and contributing to pro-inflammatory states, which hold significance for diagnostic and therapeutic applications.

Methods

Study participants and sample collection

The present case-control study was performed during the last quarter of 2021 at the National Institutes of Health, Islamabad, Pakistan. Adults (>18 y old) diagnosed with type 2 diabetes according to the American Diabetes Association (ADA) criteria⁶⁵ within 5 y (n = 48) and age- and sex-matched healthy controls (n = 46) were recruited via primary and occupational health-care providers to the present study. All the participants are currently residing in Islamabad and Rawalpindi, which originally belong to rural regions of Punjab and Khyber Pakhtunkhwa provinces. Exclusion criteria included type 1 diabetes, significant diseases including cardiovascular, liver, or kidney disease, malignancy, bariatric, or any major surgical procedure in the previous 3 months, gastrointestinal diseases or symptoms of constipation or diarrhea, pregnancy or breastfeeding, use of antibiotics in the previous 3 months, being underweight (BMI < 18.5 kg/m²), and alcohol consumption.

All participants underwent anthropometric measurements at the clinic, and medical/drug history was documented for individuals with T2D. All participants also completed simplified

food frequency questionnaires. The questionnaires were further converted to four types of diet for microbiota analysis (Table S5); this reductionist approach was mainly for convenience, but also recent studies suggest that dietary patterns and food groups associate more strongly with microbiota composition than conventional nutrients.⁶⁶ Participants were informed of the need for fresh samples, and thus stool samples were collected in sterile containers provided to the participants and immediately transferred to -80°C freezer within 1 h of defecation. Blood samples were analyzed for HbA1c and for lipid profiles (total cholesterol, triglycerides, LDL, and HDL) by an automated analyzer (Cobas Integra 700; Hoffman-La Roche, Basel, Switzerland). The rest of the plasma samples were stored at -20°C until further analysis.

The healthy Finnish controls were previously recruited as part of a dietary intervention trial investigating the effects of partly replacing animal proteins with plant proteins on health; the inclusion and exclusion criteria have been described elsewhere.⁶⁷ The fecal samples collected at baseline from 50 Finnish adults that were age-, sex- and BMI-matched with the Pakistani controls were included in the present study.

The study was carried out in accordance with the Declaration of Helsinki. The protocols were approved by the Ethics Review Committee (ERC) of the National Institutes of Health, Islamabad, and the Medical Ethical Committees of the Hospital District of Helsinki and Uusimaa and HUCH. All subjects gave written informed consents.

Publicly available datasets from other cohorts

To compare the Pakistan gut microbiota with that of other cohorts or populations, we systemically searched for publicly available 16S rRNA gene amplicon (targeting the V3-V4 region) datasets of T2D case-control studies with available metadata or identification of T2D status. Sequences from five published studies were subsequently downloaded from the National Center for Biotechnology Information and the DNA Data Bank of Japan: PRJNA661673 (China),⁶⁸ PRJNA766337 (Japan),⁶⁹ PRJDB9293 (Indonesia),⁷⁰ PRJNA588353 (Sudan),⁷¹ and

PRJNA554535 (Pakistan; an independent cohort consisting of participants residing in Islamabad).¹⁵ An endogamous agriculturist Indian cohort of healthy adults (PRJNA399246) was also included,⁷² representative of a non-industrialized population from the same geographic region. The characteristics of the included cohorts are summarized in Table S1.

DNA extraction and sequencing of 16S rRNA gene and internal transcribed spacer-1 (ITS-1) amplicons

DNA was extracted from the Pakistani and Finnish participants' fecal samples using the same procedure, Repeated Bead Beating (RBB) method,⁷³ with the following modifications for automated DNA purification: Approximately, 0.25 g of fecal samples and 340 μL and 145 μL of lysis buffer were used on the first and second rounds of bead beating, respectively. Then, 200 μL of the clarified supernatant collected from the two bead beating rounds was used for DNA extraction with the Ambion Magmax™ -96 DNA Multi-Sample Kit (4413022, Thermo Fisher Scientific, USA) using the KingFisher™ Flex automated purification system (ThermoFisher Scientific, USA). DNA was quantified using Quanti-iT™ Pico Green dsDNA Assay (Invitrogen, San Diego, CA, USA). Library preparation and Illumina MiSeq sequencing of the hypervariable V3-V4 regions of the 16S rRNA gene using primers 341 F/785 R were performed as previously described.⁷⁴ To characterize the gut fungal community, we performed ITS-1 sequencing for the Pakistani samples using a two-step PCR protocol described in detail elsewhere.⁷⁵ PCR-amplicons of the ITS-1 region were generated using ITS1F and ITS2 primers.^{75,76}

Quantification of butyrate production capacity and eubacterial quantitative PCR (qPCR)

The butyryl-CoA:acetate CoA-transferase gene and total bacterial abundance were quantified using fecal DNA with the degenerate primers BCoATscrF/R and the universal primers 331 F/797 R by qPCR, respectively. The qPCR assays have been described in detail previously^{74,77} and were performed in triplicate on a BioRad C1000 Touch thermal cycler (BioRad, Hercules, CA) with HOT FIREPol® EvaGreen® qPCR Mix Plus (Solis

BioDyne, Tartu, Estonia). For quantification of the butyryl-CoA:acetate CoA-transferase gene, the mean threshold cycle (Ct) per sample (after excluding triplicates with Ct values that differed >0.5) was used as a proxy for the abundance of the target gene. For quantification of total eubacteria, the 10-log-fold standard curves ranging from 10^2 to 10^7 copies were produced using full-length amplicons of 16S rRNA gene of *Bifidobacterium longum* to convert the threshold cycle (Ct) values into the average estimates of target bacterial genomes present in 1 g of feces (copy numbers/g of wet feces) in the assays.⁷⁴

Data processing and statistical analysis

For the bacterial microbiota, demultiplexed reads after adaptor removal were processed using DADA2⁷⁸ to generate amplicon sequence variants (ASVs). Taxonomic classification was performed using a naive Bayes classifier against the SILVA 132 reference database⁷⁸ for all the included cohorts. Species assignment was performed using DADA2 by exact string matching against the SILVA species assignment training database.⁷⁹ The 16S rRNA gene sequences for the Pakistani and Finnish samples from this study have been deposited in the European Nucleotide Archive (<https://www.ebi.ac.uk/ena>) under accession numbers PRJEB53017 and PRJEB53018, respectively. For the fungal microbiota, the ITS-1 sequences were pre-processed following the DADA2 ITS Pipeline Workflow (1.8) available on the official DADA2 homepage, and taxonomically annotated using UNIITE ver. 7.2 Dynamic Classifier as the reference database with the similarity threshold at 0.97.⁸⁰ The ITS-1 sequences have been deposited in the European Nucleotide Archive (<https://www.ebi.ac.uk/ena>) under the accession number PRJEB53019.

To infer the functional contribution of bacterial communities from 16S rRNA gene sequencing data, metagenome prediction was carried out using PICRUSt2 (Phylogenetic Investigation of Communities by Reconstruction of Unobserved States)⁸¹ evaluating KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways.⁸²

Differential abundance for bacterial and fungal taxa or KEGG pathways between case and control

participants was identified with the *DESeq2* package.⁸³ *DESeq2* employs a generalized linear model of counts based on a negative binomial distribution, scaled by a normalization factor that accounts for differences in sequencing depth between samples. Significance testing was then assessed using the Wald test. Non-count variables (anthropometric, biochemical, and other measurements) were analyzed with the Wilcoxon signed-rank test, t-test, or chi-square test depending on the data distribution.

Microbiota α -diversity (observed richness and Shannon diversity index) was estimated using the *vegan* package.⁸⁴ Overall microbiota structure was assessed by principal coordinate analysis (PCoA) based on β -diversity computed using the Bray-Curtis dissimilarity matrix, representing the compositional dissimilarity between samples or groups. Significant differences between groups were tested using nonparametric multivariate analysis of variance (PERMANOVA).⁸⁴ The associations between continuous or categorical variables and β -diversity were calculated using the *envfit* function in the *vegan* package,⁸⁴ and P values were determined using 999 permutations. Associations between relative abundances of bacterial taxa and other measurements were assessed using Spearman's correlation and visualized by the *R* package *circlize*.⁸⁵ The *R* package *Simpsons* was additionally employed to ensure the absence of Simpson's paradox that misleads the interpretation of correlative analysis.⁸⁶

For microbiota-based T2D classification, a random forest model (100 random forest replicates of 1000 trees using the default settings of the *R* package *randomForest*) was built based on a list of prevalent bacterial genera (prevalence >10%) detected in our Pakistani cohort (Table S3). The model performance was evaluated using repeated k-fold cross-validation (fivefold, 10 repetitions) implemented in the *caret* package⁸⁷ and scored the predictive power in a receiver operating characteristic (ROC) analysis. The top 10 important input genera were subsequently ranked according to the mean decrease in Gini provided by the *R* package *randomForest*. To assess how well the classifier trained on one cohort can be generalized to the other cohorts, study-to-study model transfer was performed as described previously.²⁷ Briefly,

one random forest model was built based on the aforementioned list of input genera for each study/cohort and then applied to the other studies using the same parameters to generate area under curves (AUCs) for cross-applications.

Statistical analyses were performed with the statistical program R version 3.5.0 and RStudio version 0.99.903. P values were corrected for multiple comparisons by using the Benjamini–Hochberg procedure (FDR). P values and FDR-adjusted P-values <0.05 were considered significant.

Inference of clade of *Prevotella copri* ASVs

The reconstructed genome sequences of *Prevotella copri* by Tett *et al.*¹⁹ were collected and converted into a BLAST database using the BLAST+ package.⁸⁸ The sequence of each ASV annotated as *Prevotella copri* was queried to the database using the *blastn* command, and the reconstructed genomes containing the ASV sequence (100% identify) were retrieved. The clades of these genomes were assigned as the clade of the ASV.

Measurement of Toll-like receptor (TLR) activation and lipopolysaccharide binding protein (LBP)

To quantify potential microbial products in circulation, TLR activation capacity and levels of LBP in plasma samples were used as a proxy. The protocol using HEK-Blue™-hTLR2 and HEK-Blue™-hTLR4 reporter cell lines expressing human TLR2 and TLR4 (InvivoGen, San Diego, CA) was employed according to the manufacturer's instructions. The HEK-Blue™ hTLR cells were grown for two passages with medium supplemented without selective antibiotics provided by the manufacturer and then passaged in medium with selective antibiotics that was also used for the experiment. The assay was performed when the cells were in passage 10–15 by adding approximately 10⁵ HEK-Blue™-hTLR2 cells and 1.4 × 10⁵ HEK-Blue™-hTLR4 cells in 96-well plates containing 20 µl of plasma samples and incubated for 24 hours at 37°C under an atmosphere of 5% CO₂/95% air. Twenty microliters of the cell culture supernatants were added to 180 µl of the QUANTI-Blue substrate in a 96-well plate. The mixtures were then incubated at 37°C for 3 hours

and secreted embryonic alkaline phosphatase levels were determined using a spectrophotometer at 630 nm. Lipoteichoic acid (Invivogen, LTA-BS; 1 µg/ml, 500 ng/ml, and 100 ng/ml), peptidoglycan (Invivogen, PGN-BS; 0.1 µg/ml, 5 µg/ml, 10 µg/ml) and lipopolysaccharide (Sigma Aldrich, LPS, 1000 ng/ml, 100 ng/ml, 10 ng/ml) were used as positive controls, and cell culture medium served as a negative control. Due to reagent availability, the TLR2 experiment was performed for the samples derived from 50% participants whose age, sex, and BMI matched the entire cohort (control = 23, T2D = 24). Plasma LBP was measured by quantitative ELISA using human LBP DuoSet kits (R&D Systems, Minneapolis, MN) according to the manufacturer's instructions.

Authors' contributions

Afshan Saleem: Conceptualization, Methodology, Formal analysis, Investigation, Resources, Data Curation, Project administration, Writing – Original Draft, Visualization, Funding acquisition

Aamer Ikram: Conceptualization, Resources, Project administration, Writing – Review & Editing

Evgenia Dikareva: Methodology, Investigation, Writing – Review & Editing

Emilia Lahtinen: Methodology, Investigation, Writing – Review & Editing

Dollwin Matharu: Methodology, Investigation, Writing – Review & Editing, Supervision

Anne-Maria Pajari: Resources, Data Curation, Writing – Review & Editing

Willem M. de Vos: Methodology, Resources, Funding acquisition, Writing – Review & Editing

Fariha Hasan: Conceptualization, Resources, Project administration, Writing – Review & Editing

Anne Salonen: Conceptualization, Methodology, Resources, Writing – Review & Editing, Supervision, Funding acquisition

Ching Jian: Conceptualization, Methodology, Formal analysis, Investigation, Data Curation, Writing – Original Draft, Visualization, Supervision

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This study was supported by the Higher Education Commission, Pakistan, under the International Research Support Initiative Program (Afshan Saleem), the Research

Programs Unit of the Faculty of Medicine, the University of Helsinki (Anne Salonen) and Advanced Research Grant 250172 (MicrobesInside) of the European Research Council (Willem M. de Vos). The open access was funded by Helsinki University Library.

ORCID

Anne Salonen  <http://orcid.org/0000-0002-6960-7447>

Ching Jian  <http://orcid.org/0000-0003-0577-8834>

Data availability

The datasets generated in this study are available in the European Nucleotide Archive (ENA) repository, under accession no. PRJEB53017, PRJEB53018, and PRJEB53019. The associated metadata is available in Table S4. Supplemental data for this article can be accessed on publisher's website.

References

- de Vos WM, Tilg H, Van Hul M, Cani PD. Gut microbiome and health: mechanistic insights. *Gut*. 2022;71(5):1020–1032. doi:10.1136/gutjnl-2021-326789.
- Sonnenburg JL, Sonnenburg ED. Vulnerability of the industrialized microbiota. *Science*. 2019;366. doi:10.1126/science.aaw9255.
- Abdill RJ, Adamowicz EM, Blekhman R. Public human microbiome data are dominated by highly developed countries. *PLoS Biol*. 2022;20:e3001536. doi:10.1371/journal.pbio.3001536.
- Almeida A, Mitchell AL, Boland M, Forster SC, Gloor GB, Tarkowska A, Lawley TD, Finn RD. A new genomic blueprint of the human gut microbiota. *Nature*. 2019;568:499–504. doi:10.1038/s41586-019-0965-1.
- Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, Beghini F, Manghi P, Tett A, Ghensi P, et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell*. 2019;176:649–62.e20. doi:10.1016/j.cell.2019.01.001.
- Lu J, Zhang L, Zhai Q, Zhao J, Zhang H, Lee YK, Lu W, Li M, Chen W. Chinese gut microbiota and its associations with staple food type, ethnicity, and urbanization. *NPJ Biofilms Microbiomes*. 2021;7:71. doi:10.1038/s41522-021-00245-0.
- He Y, Wu W, Zheng HM, Li P, McDonald D, Sheng HF, Chen MX, Chen ZH, Ji GY, Zheng ZD, et al. Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nat Med*. 2018;24:1532–1535. doi:10.1038/s41591-018-0164-x.
- Deschasaux M, Bouter KE, Prodan A, Levin E, Groen AK, Herrema H, Tremaroli V, Bakker GJ, Attaye I, Pinto-Sietsma S-J, et al. Depicting the composition of gut microbiota in a population with varied ethnic origins but shared geography. *Nat Med*. 2018;24(10):1526–1531. doi:10.1038/s41591-018-0160-1.
- Brooks AW, Priya S, Blekhman R, Bordenstein SR. Gut microbiota diversity across ethnicities in the United States. *PLoS Biol*. 2018;16:e2006842. doi:10.1371/journal.pbio.2006842.
- Ang QY, Alba DL, Upadhyay V, Bisanz JE, Cai J, Lee HL, Barajas E, Wei G, Noecker C, Patterson AD, et al. The East Asian gut microbiome is distinct from colocalized White subjects and connected to metabolic health. *eLife*. 2021;10. doi:10.7554/eLife.70349.
- Alvarez-Silva C, Kashani A, Hansen TH, Pinna NK, Anjana RM, Dutta A, Saxena S, Støy J, Kampmann U, Nielsen T, et al. Trans-ethnic gut microbiota signatures of type 2 diabetes in Denmark and India. *Genome Med*. 2021;13(1):37. doi:10.1186/s13073-021-00856-4.
- Hermes GDA, Reijnders D, Kootte RS, Goossens GH, Smidt H, Nieuwdorp M, Blaak EE, Zoetendal EG. Individual and cohort-specific gut microbiota patterns associated with tissue-specific insulin sensitivity in overweight and obese males. *Sci Rep*. 2020;10:7523. doi:10.1038/s41598-020-64574-4.
- Cho NH, Shaw JE, Karuranga S, Huang Y, da Rocha Fernandes JD, Ohlrogge AW, Malanda B. IDF Diabetes Atlas: global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Res Clin Pract*. 2018;138:271–281. doi:10.1016/j.diabres.2018.02.023.
- Batool M, Ali SB, Jaan A, Khalid K, Ali SA, Kamal K, Raja AA, AA F, Nasir A. Initial sequencing and characterization of the gastrointestinal and oral microbiota in urban Pakistani adults. *Front Cell Infect Microbiol*. 2020;10:409. doi:10.3389/fcimb.2020.00409.
- Ahmad A, Yang W, Chen G, Shafiq M, Javed S, Ali Zaidi SS, Shahid R, Liu C, Bokhari H. Analysis of gut microbiota of obese individuals with type 2 diabetes and healthy individuals. *PLoS One*. 2019;14(12):e0226372. doi:10.1371/journal.pone.0226372.
- Abellan-Schneyder I, Matchado MS, Reitmeier S, Sommer A, Sewald Z, Baumbach J, List M, Neuhaus K. Primer, pipelines, parameters: issues in 16S rRNA gene sequencing. *mSphere*. 2021;6. doi:10.1128/mSphere.01202-20.
- Rintala A, Pietilä S, Munukka E, Eerola E, Pursiheimo JP, Laiho A, Pekkala S, Huovinen P. Gut microbiota analysis results are highly dependent on the 16S rRNA gene target region, whereas the impact of DNA extraction is minor. *J Biomol Tech*. 2017;28:19–30. doi:10.7171/jbt.17-2801-003.
- Sonnenburg ED, Sonnenburg JL. The ancestral and industrialized gut microbiota and implications for human health. *Nat Rev Microbiol*. 2019;17(6):383–390. doi:10.1038/s41579-019-0191-8.
- Tett A, Huang KD, Asnicar F, Fehlner-Peach H, Pasolli E, Karcher N, Armanini F, Manghi P, Bonham K, Zolfo M, et al. The *Prevotella copri*

- complex comprises four distinct clades underrepresented in westernized populations. *Cell Host Microbe*. 2019;26(5):666–79.e7. doi:10.1016/j.chom.2019.08.018.
20. Gomez-Arango LF, Barrett HL, Wilkinson SA, Callaway LK, McIntyre HD, Morrison M, Dekker Nitert M. Low dietary fiber intake increases *Collinsella* abundance in the gut microbiota of overweight and obese pregnant women. *Gut Microbes*. 2018;9:189–201. doi:10.1080/19490976.2017.1406584.
 21. Chen J, Wright K, Davis JM, Jeraldo P, Marietta EV, Murray J, Nelson H, Matteson EL, Taneja V. An expansion of rare lineage intestinal microbes characterizes rheumatoid arthritis. *Genome Med*. 2016;8:43. doi:10.1186/s13073-016-0299-7.
 22. Lucas LN, Barrett K, Kerby RL, Zhang Q, Cattaneo LE, Stevenson D, Rey FE, Amador-Noguez D. Dominant bacterial phyla from the human gut show widespread ability to transform and conjugate bile acids. *mSystems*. 2021;e0080521. doi:10.1128/mSystems.00805-21.
 23. Gurung M, Li Z, You H, Rodrigues R, Jump DB, Morgun A, Shulzhenko N. Role of gut microbiota in type 2 diabetes pathophysiology. *EBioMedicine*. 2020;51:102590. doi:10.1016/j.ebiom.2019.11.051.
 24. Somavanshi R, Ghosh B, Sourjik V. Sugar influx sensing by the phosphotransferase system of *Escherichia coli*. *PLoS Biol*. 2016;14:e2000074. doi:10.1371/journal.pbio.2000074.
 25. Pinna NK, Anjana RM, Saxena S, Dutta A, Gnanaprakash V, Rameshkumar G, Aswath S, Raghavan S, Rani CSS, Radha V, et al. Trans-ethnic gut microbial signatures of prediabetic subjects from India and Denmark. *Genome Med*. 2021;13(1):36. doi:10.1186/s13073-021-00851-9.
 26. Cani PD, Amar J, Iglesias MA, Poggi M, Knauf C, Bastelica D, Neyrinck AM, Fava F, Tuohy KM, Chabo C, et al. Metabolic endotoxemia initiates obesity and insulin resistance. *Diabetes*. 2007;56(7):1761–1772. doi:10.2337/db06-1491.
 27. Que Y, Cao M, He J, Zhang Q, Chen Q, Yan C, Lin A, Yang L, Wu Z, Zhu D, et al. Gut bacterial characteristics of patients with type 2 diabetes mellitus and the application potential. *Front Immunol*. 2021;12:722206. doi:10.3389/fimmu.2021.722206.
 28. Forslund K, Hildebrand F, Nielsen T, Falony G, Le Chatelier E, Sunagawa S, Prifti E, Vieira-Silva S, Gudmundsdottir V, Krogh Pedersen H, et al. Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature*. 2015;528(7581):262–266. doi:10.1038/nature15766.
 29. Lahti L, Salojärvi J, Salonen A, Scheffer M, de Vos WM. Tipping elements in the human intestinal ecosystem. *Nat Commun*. 2014;5(1):4344. doi:10.1038/ncomms5344.
 30. Khan FS, Lotia-Farrukh I, Khan AJ, Siddiqui ST, Sajun SZ, Malik AA, Burfat A, Arshad MH, Codlin AJ, Reiningger BM, et al. The burden of non-communicable disease in transition communities in an Asian megacity: baseline findings from a cohort study in Karachi, Pakistan. *PLoS One*. 2013;8(2):e56008. doi:10.1371/journal.pone.0056008.
 31. Shah N, Shah Q, Shah AJ. The burden and high prevalence of hypertension in Pakistani adolescents: a meta-analysis of the published studies. *Arch Public Health*. 2018;76:20. doi:10.1186/s13690-018-0265-5.
 32. Groussin M, Poyet M, Sistiaga A, Kearney SM, Moniz K, Noel M, Hooker J, Gibbons SM, Segurel L, Froment A, et al. Elevated rates of horizontal gene transfer in the industrialized human microbiome. *Cell*. 2021;184:2053–67.e18. doi:10.1016/j.cell.2021.02.052.
 33. Zuo T, Kamm MA, Colomel JF, Ng SC. Urbanization and the gut microbiota in health and inflammatory bowel disease. *Nat Rev Gastroenterol Hepatol*. 2018;15:440–452. doi:10.1038/s41575-018-0003-z.
 34. Sonnenburg ED, Smits SA, Tikhonov M, Higginbottom SK, Wingreen NS, Sonnenburg JL. Diet-induced extinctions in the gut microbiota compound over generations. *Nature*. 2016;529:212–215. doi:10.1038/nature16504.
 35. Keohane DM, Ghosh TS, Jeffery IB, Molloy MG, O'Toole PW, Shanahan F. Microbiome and health implications for ethnic minorities after enforced lifestyle changes. *Nat Med*. 2020;26:1089–1095. doi:10.1038/s41591-020-0963-8.
 36. Ayeni FA, Biagi E, Rampelli S, Fiori J, Soverini M, Audu HJ, Cristino S, Caporali L, Schnorr SL, Carelli V, et al. Infant and adult gut microbiome and metabolome in rural Bassa and urban settlers from Nigeria. *Cell Rep*. 2018;23(10):3056–3067. doi:10.1016/j.celrep.2018.05.018.
 37. Vangay P, Johnson AJ, Ward TL, Al-Ghalith GA, Shields-Cutler RR, Hillmann BM, Lucas SK, Beura LK, Thompson EA, Till LM, et al. US immigration westernizes the human gut microbiome. *Cell*. 2018;175:962–72.e10. doi:10.1016/j.cell.2018.10.029.
 38. Martínez I, Stegen JC, Maldonado-Gómez MX, Eren AM, Siba PM, Greenhill AR, Walter J. The gut microbiota of rural Papua New Guineans: composition, diversity patterns, and ecological processes. *Cell Rep*. 2015;11:527–538. doi:10.1016/j.celrep.2015.03.049.
 39. Hansen MEB, Rubel MA, Bailey AG, Ranciaro A, Thompson SR, Campbell MC, Beggs W, Dave JR, Mokone GG, Mpoloka SW, et al. Population structure of human gut bacteria in a diverse cohort from rural Tanzania and Botswana. *Genome Biol*. 2019;20:16. doi:10.1186/s13059-018-1616-9.
 40. Sun S, Wang H, Howard AG, Zhang J, Su C, Wang Z, Du S, Fodor AA, Gordon-Larsen P, Zhang B, et al. Loss of novel diversity in human gut microbiota associated with ongoing urbanization in China. *mSystems*. 2022;7(4):e0020022. doi:10.1128/msystems.00200-22.
 41. Micha R, Khatibzadeh S, Shi P, Andrews KG, Engell RE, Mozaffarian D. Global, regional and national consumption of major food groups in 1990 and 2010: a systematic analysis including 266 country-specific nutrition surveys worldwide. *BMJ Open*. 2015;5(9):e008705. doi:10.1136/bmjopen-2015-008705.

42. de Goffau MC, Jallow AT, Sanyang C, Prentice AM, Meagher N, Price DJ, Reville PA, Parkhill J, Pereira DIA, Wagner J, et al. Gut microbiomes from Gambian infants reveal the development of a non-industrialized Prevotella-based trophic network. *Nat Microbiol.* 2022;7:132–144. doi:10.1038/s41564-021-01023-6.
43. Sarwar T. Infant feeding practices of Pakistani mothers in England and Pakistan. *J Hum Nutr Diet.* 2002;15:419–428. doi:10.1046/j.1365-277x.2002.00395.x.
44. Manikam L, Sharmila A, Dharmaratnam A, Alexander EC, Kuah JY, Prasad A, Ahmed S, Lingam R, Lakhanpaul M. Systematic review of infant and young child complementary feeding practices in South Asian families: the Pakistan perspective. *Public Health Nutr.* 2018;21:655–668. doi:10.1017/s1368980017002956.
45. Zhang J, Huang YJ, Yoon JY, Kemmitt J, Wright C, Schneider K, Sphabmixay P, Hernandez-Gordillo V, Holcomb SJ, Bhushan B, et al. Primary human colonic mucosal barrier crosstalk with super oxygen-sensitive *Faecalibacterium prausnitzii* in continuous culture. *Med.* 2021;2:74–98.e9. doi:10.1016/j.medj.2020.07.001.
46. Bilen M, Cadoret F, Richez M, Tomei E, Daoud Z, Raoult D, Fournier PE. *Libanicoccus massiliensis* gen. nov., sp. nov., a new bacterium isolated from human stool. *New Microbes New Infect.* 2018;21:63–71. doi:10.1016/j.nmni.2017.11.001.
47. Sakamoto M, Ikeyama N, Murakami T, Mori H, Yuki M, Ohkuma M. Comparative genomics of *Parolsenella catena* and *Libanicoccus massiliensis*: reclassification of *Libanicoccus massiliensis* as *Parolsenella massiliensis* comb. nov. *Int J Syst Evol Microbiol.* 2019;69:1123–1129. doi:10.1099/ijsem.0.003283.
48. Perez-Muñoz ME, Sugden S, Harmsen HJM, T Hart BA, Laman JD, Walter J. Nutritional and ecological perspectives of the interrelationships between diet and the gut microbiome in multiple sclerosis: insights from marmosets. *iScience.* 2021;24:102709. doi:10.1016/j.isci.2021.102709.
49. Zhu S, Xu K, Jiang Y, Zhu C, Suo C, Cui M, Wang Y, Yuan Z, Xue J, Wang J, et al. The gut microbiome in subclinical atherosclerosis: a population-based multi-phenotype analysis. *Rheumatology.* 2021;61(1):258–269. doi:10.1093/rheumatology/keab309.
50. Giliberti R, Cavaliere S, Mauriello IE, Ercolini D, Pasolli E. Host phenotype classification from human microbiome data is mainly driven by the presence of microbial taxa. *PLoS Comput Biol.* 2022;18:e1010066. doi:10.1371/journal.pcbi.1010066.
51. Wirbel J, Zych K, Essex M, Karcher N, Kartal E, Salazar G, Bork P, Sunagawa S, Zeller G. Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox. *Genome Biol.* 2021;22:93. doi:10.1186/s13059-021-02306-1.
52. Leite AZ, Rodrigues NC, Gonzaga MI, Paiolo JCC, de Souza CA, Stefanutto NAV, Omori WP, Pinheiro DG, Brisotti JL, Matheucci Junior E, et al. Detection of increased plasma interleukin-6 levels and prevalence of *Prevotella copri* and *Bacteroides vulgatus* in the feces of type 2 diabetes patients. *Front Immunol.* 2017;8:1107. doi:10.3389/fimmu.2017.01107.
53. Doumatey AP, Adeyemo A, Zhou J, Lei L, Adebamowo SN, Adebamowo C, Rotimi CN. Gut microbiome profiles are associated with type 2 diabetes in urban Africans. *Front Cell Infect Microbiol.* 2020;10:63. doi:10.3389/fcimb.2020.00063.
54. Tett A, Pasolli E, Masetti G, Ercolini D, Segata N. Prevotella diversity, niches and interactions with the human host. *Nat Rev Microbiol.* 2021;19:585–599. doi:10.1038/s41579-021-00559-y.
55. Schnorr SL, Candela M, Rampelli S, Centanni M, Consolandi C, Basaglia G, Turroni S, Biagi E, Peano C, Severgnini M, et al. Gut microbiome of the Hadza hunter-gatherers. *Nat Commun.* 2014;5:3654. doi:10.1038/ncomms4654.
56. Gomes JMG, Costa JA, Alfenas RCG. Metabolic endotoxemia and diabetes mellitus: a systematic review. *Metabolism.* 2017;68:133–144. doi:10.1016/j.metabol.2016.12.009.
57. Novitsky TJ. Limitations of the *Limulus* amoebocyte lysate test in demonstrating circulating lipopolysaccharides. *Ann N Y Acad Sci.* 1998;851:416–421. doi:10.1111/j.1749-6632.1998.tb09018.x.
58. Munford RS. Endotoxemia-menace, marker, or mistake? *J Leukoc Biol.* 2016;100:687–698. doi:10.1189/jlb.3RU0316-151R.
59. Molteni M, Gemma S, Rossetti C. The role of toll-like receptor 4 in infectious and noninfectious inflammation. *Mediators Inflamm.* 2016;2016:6978936. doi:10.1155/2016/6978936.
60. Jurburg SD, Konzack M, Eisenhauer N, Heintz-Buschart A. The archives are half-empty: an assessment of the availability of microbial community sequencing data. *Commun Biol.* 2020;3:474. doi:10.1038/s42003-020-01204-9.
61. Lee CB, Chae SU, Jo SJ, Jerng UM, Bae SK. The relationship between the gut microbiome and metformin as a key for treating type 2 diabetes mellitus. *Int J Mol Sci.* 2021;22. doi:10.3390/ijms22073566.
62. Elbere I, Silamikelis I, Dindune K, Ustinova I, Zaharenko M, Silamikele L, Rovite V, Gudra D, Konrade I. Baseline gut microbiome composition predicts metformin therapy short-term efficacy in newly diagnosed type 2 diabetes patients. *PLoS One.* 2020;15:e0241338. doi:10.1371/journal.pone.0241338.
63. Wu H, Esteve E, Tremaroli V, Khan MT, Caesar R, Mannerås-Holm L, Ståhlman M, Olsson LM, Serino M, Planas-Félix M, et al. Metformin alters the gut microbiome of individuals with treatment-naïve type 2 diabetes, contributing to the therapeutic effects of the drug. *Nat Med.* 2017;23(7):850–858. doi:10.1038/nm.4345.
64. Wagner Mackenzie B, Waite DW, Taylor MW. Evaluating variation in human gut microbiota profiles due to DNA extraction method and inter-subject differences. *Front Microbiol.* 2015;6:130. doi:10.3389/fmicb.2015.00130.

65. American Diabetes Association. Classification and diagnosis of diabetes: standards of medical care in diabetes-2020. *Diabetes Care*. 2020;43:S14–s31. doi:10.2337/dc20-S002.
66. Choi Y, Hoops SL, Thoma CJ, Johnson AJ. A guide to dietary pattern-microbiome data integration. *J Nutr*. 2022;152:1187–1199. doi:10.1093/jn/nxac033.
67. Päivärinta E, Itkonen ST, Pellinen T, Lehtovirta M, Erkkola M, Pajari AM. Replacing animal-based proteins with plant-based proteins changes the composition of a whole Nordic diet – a randomised clinical trial in healthy Finnish adults. *Nutrients*. 2020;12. doi:10.3390/nu12040943.
68. Zhang Z, Tian T, Chen Z, Liu L, Luo T, Dai J. Characteristics of the gut microbiome in patients with prediabetes and type 2 diabetes. *PeerJ*. 2021;9:e10952. doi:10.7717/peerj.10952.
69. Kondo Y, Hashimoto Y, Hamaguchi M, Ando S, Kaji A, Sakai R, Inoue R, Kashiwagi S, Mizushima K, Uchiyama K, et al. Unique habitual food intakes in the gut microbiota cluster associated with type 2 diabetes mellitus. *Nutrients*. 2021;13. doi:10.3390/nu13113816.
70. Therdtatha P, Song Y, Tanaka M, Mariyatun M, Almunifah M, Manurung NEP, Indriarsih S, Lu Y, Nagata K, Fukami K, et al. Gut microbiome of Indonesian adults associated with obesity and type 2 diabetes: a cross-sectional study in an Asian City, Yogyakarta. *Microorganisms*. 2021;9(5):897. doi:10.3390/microorganisms9050897.
71. Almugadam BS, Liu Y, Chen SM, Wang CH, Shao CY, Ren BW, Tang L. Alterations of gut microbiota in type 2 diabetes individuals and the confounding effect of anti-diabetic agents. *J Diabetes Res*. 2020;2020:7253978. doi:10.1155/2020/7253978.
72. Chaudhari DS, Dhotre DP, Agarwal DM, Gaike AH, Bhalerao D, Jadhav P, Mongad D, Lubree H, Sinkar VP, Patil UK, et al. Gut, oral and skin microbiome of Indian patrilineal families reveal perceptible association with age. *Sci Rep*. 2020;10(1):5685. doi:10.1038/s41598-020-62195-5.
73. Salonen A, Nikkila J, Jalanka-Tuovinen J, Immonen O, Rajilic-Stojanovic M, Kekkonen RA, Palva A, de Vos WM. Comparative analysis of fecal DNA extraction methods with phylogenetic microarray: effective recovery of bacterial and archaeal DNA using mechanical cell lysis. *J Microbiol Methods*. 2010;81:127–134. doi:10.1016/j.mimet.2010.02.007.
74. Jian C, Luukkonen P, Yki-Järvinen H, Salonen A, Korpela K. Quantitative PCR provides a simple and accessible method for quantitative microbiota profiling. *PLoS One*. 2020;15:e0227285. doi:10.1371/journal.pone.0227285.
75. Virtanen SS, Kanerva S, Nieminen S, Kalliala P, Salonen I. Metagenome-validated parallel amplicon sequencing and text mining-based annotations for simultaneous profiling of bacteria and fungi: vaginal microbiome and mycobiota in healthy women. 2021. Preprint (Version 1) available at Research Square. doi:10.21203/rs.3.rs-321778/v1.
76. Gardes M, Bruns TD. ITS primers with enhanced specificity for basidiomycetes—application to the identification of mycorrhizae and rusts. *Mol Ecol*. 1993;2:113–118. doi:10.1111/j.1365-294x.1993.tb00005.x.
77. Louis P, Flint HJ. Development of a semiquantitative degenerate real-time PCR-based assay for estimation of numbers of butyryl-coenzyme A (CoA) CoA transferase genes in complex bacterial samples. *Appl Environ Microbiol*. 2007;73:2009–2012. doi:10.1128/aem.02561-06.
78. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods*. 2016;13:581–583. doi:10.1038/nmeth.3869.
79. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res*. 2013;41:D590–6. doi:10.1093/nar/gks1219.
80. Nilsson RH, Larsson KH, Taylor AFS, Bengtsson-Palme J, Jeppesen TS, Schigel D, Kennedy P, Picard K, Glöckner FO, Tederso L, et al. The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Res*. 2019;47:D259–d64. doi:10.1093/nar/gky1022.
81. Douglas GM, Maffei VJ, Zaneveld JR, Yurgel SN, Brown JR, Taylor CM, Huttenhower C, Langille MGI. PICRUSt2 for prediction of metagenome functions. *Nat Biotechnol*. 2020;38:685–688. doi:10.1038/s41587-020-0548-6.
82. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res*. 2017;45:D353–d61. doi:10.1093/nar/gkw1092.
83. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15:550. doi:10.1186/s13059-014-0550-8.
84. Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'hara R, Simpson GL, Solymos P, Stevens MHH, Wagner H, et al. Package 'vegan'. *Community Ecology Package*, Version. 2013;2.
85. Gu Z, Gu L, Eils R, Schlesner M, Brors B. circlize Implements and enhances circular visualization in R. *Bioinformatics*. 2014;30:2811–2812. doi:10.1093/bioinformatics/btu393.
86. Kievit RA, Frankenhuis WE, Waldorp LJ, Borsboom D. Simpson's paradox in psychological science: a practical guide. *Front Psychol*. 2013;4:513. doi:10.3389/fpsyg.2013.00513.
87. Kuhn M. Building Predictive Models in R Using the caret Package. *J Statistical Software*. 2008;28. doi:10.18637/jss.v028.i05.
88. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. *BMC Bioinform*. 2009;10:421. doi:10.1186/1471-2105-10-421.