

RESEARCH PAPER

 OPEN ACCESS 

Metagenomic analysis reveals distinct patterns of gut lactobacillus prevalence, abundance, and geographical variation in health and disease

Tarini Shankar Ghosh ^{a*}, Jerome Arnoux^{a,b*}, and Paul W. O'Toole ^a

^aDepartment: School of Microbiology and APC Microbiome Ireland Institution, University College Cork, Cork, Ireland; ^bDepartment: UFR des Sciences et Techniques Institution, Université De Rouen, Normandie, France

ABSTRACT

Lactobacilli are exploited extensively for food fermentation and biotechnology. Some food and gut isolates have been developed as probiotics, for which species that may be commensal to the human host are considered desirable. However, the robustness of defining original niches for lactobacilli – food, environment, the gut – is questionable, and culture-independent analyses of prevalence in different human populations is lacking. Here we analyzed the abundance of lactobacilli in 6,154 subjects from a database of highly curated fecal shotgun metagenomics data spanning 25 nationalities, with ages ranging from infancy to 102 years. Twenty-five species were detected, which we assigned into low, medium, and high prevalence groups. The microbiome of apparently healthy individuals could be categorized into 6 clusters or Lactobacillotypes (LbTypes), with three of the Lbtypes being dominated by *L. delbrueckii*, *L. ruminis*, *L. casei*, and the other three comprising a combination of different species. These Lactobacillus clusters exhibit distinct global abundance patterns. The cluster prevalences also display distinct age-specific trends influenced by geography, with overall lactobacillus prevalence increasing significantly with age in North America and Europe but declining with age in non-Westernized societies. Regression analysis stratified by regional location identified distinct associations of the Lactobacillotypes with age, BMI, and gender. Cirrhosis, fatty-liver, IBD and T2D were characterized by net gain of lactobacilli, whereas hypertension patients harbored depleted lactobacillus levels. Collectively these data indicate that the species abundance of gut lactobacilli is moderated by geography, diet, and interaction with the whole microbiome, and has strong interactions with diseases associated with a western lifestyle.

ARTICLE HISTORY

Received 10 April 2020
Revised 22 August 2020
Accepted 28 August 2020

KEYWORDS

Gut microbiome;
lactobacillus; regional
variations; age variations;
probiotic



Introduction

Humans, mammals, insects and plants harbor distinct communities of microorganisms with whom they have co-evolved.^{1–3} Despite the challenges of defining the “normal” or health-associated state of the microbiome,⁴ there is emerging consensus that alterations in the composition and function of the animal gut microbiome are associated with pathophysiological syndromes or disease, both intestinal and extraintestinal.⁵ The precise molecular mechanisms whereby gut microbes could be involved in disease are still largely unexplained, but they include effects on metabolism,⁶ immunity/inflammation,⁷ tumorigenesis,⁸ and signaling.⁹ Analysis of the microbiome state may also be informative for assessing risk of, diagnosing or managing disease.^{10–13}


The genus *Lactobacillus* encompasses an unusually diverse number of species that share the

property of being found in nutrient-rich environments.¹⁴ Lactobacilli have been exploited extensively for food preservation,¹⁵ for biotechnological applications,¹⁶ and as health-promoting “probiotics”.¹⁷ The phenotypic diversity of the genus *Lactobacillus* is reflected in extraordinarily high genomic diversity, approaching that of other bacterial families.^{18,19} The isolation sources for most lactobacilli may be broadly categorized as humans, animals, plants, food and environment, and major “lifestyle” assignment groupings coincide remarkably with phylogenomic clades,²⁰ indicating concerted evolution for niche adaptation.

Specialization of some lactobacillus species toward the human gut could indicate a commensal role, so it has historically been of interest to identify such species. Culture from human postmortem intestinal biopsies identified

CONTACT Paul W. O'Toole  pwotoole@ucc.ie  School of Microbiology and APC Microbiome Ireland, Cork, Ireland

*These authors contributed equally to this work

 Supplemental data for this article can be accessed on the [publisher's website](#).

© 2020 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.

L. gasseri, *L. reuteri*, *L. salivarius*, *L. casei*, *L. plantarum* and *L. buchneri* as the most common lactobacillus species,²¹ and lactobacilli were considered until relatively recently to be dominant taxa in the normal microbiota that reached greatest numbers in the small bowel.²⁰ However, data from an early tranche of molecular studies reviewed by Walter,²² revealed low abundance levels (i.e., less than 1.0% total abundance) of sequences related to lactobacilli in/on fecal material or intestinal biopsies. The species most commonly found in the human gut, in addition to those listed above from the review by Reuter,²¹ include *L. acidophilus*, *L. crispatus*, *L. johnsonii*, *L. ruminis*, *L. casei/paracasei*, *L. rhamnosus*, *L. plantarum*, *L. fermentum*, *L. brevis*, *L. delbrueckii*, *L. sakei*, *L. vaginalis*, and *L. curvatus*.²² Identifying which lactobacilli are autochthonous to the human gut (formed where found) has also been of interest for the development of probiotics, based on an early misconception that species commonly used as probiotics must necessarily be able to transit the intestinal tract and then colonize the gut.²³ In reality, commensalism or autochthony is not an *a priori* requirement for an ingested microbe to have a beneficial effect on the host, and because so many lactobacillus species are naturally found in raw or fermented foods, isolating a given species from human stool does not reliably guarantee that species is autochthonous. Finally, many of the older culture-based literature relied on ambiguous phenotypic traits or since-altered species assignments,²⁴ making it hard to combine published datasets reliably to generate larger subject numbers for determining prevalence.

When administered as probiotics to humans, selected species and strains of lactobacilli have conferred benefits including alleviation of infant colic,²⁵ amelioration of lactose intolerance²⁶ and reduced symptoms of atopic dermatitis.²⁷ Many lactobacilli are present in fermented foods that themselves have reported health benefits.^{28,29} However, although human gut lactobacilli are generally considered beneficial, there are some reports of elevated lactobacillus abundance in the microbiome of people with some diseases e.g. Type 2 Diabetes,³⁰ obesity,³¹ liver cirrhosis³² and even systemic autoimmunity.³³ However, these findings are controversial and may indicate association with disease symptoms rather than causation, because other

studies have even suggested beneficial effects in these diseases.^{34–38} Lactobacilli occasionally cause bacteremia or sepsis, almost always in immunocompromised patients,³⁹ and sometimes with strains administered as probiotics rather than strains already present in the subject.^{40,41}

The overall composition and function of the gut microbiome is influenced by external factors including geographical region,⁴² ethnicity,⁴³ and diet.^{44,45} How these factors intersect to modulate the lactobacillus species that are prevalent or abundant on a global basis is currently unknown. The availability of a large number of metagenomic datasets from globally distributed cohorts, apparently healthy controls and case studies, has allowed us to dissect the interaction of location, age, health, and disease with the abundance of commensal lactobacilli. We observe that the Lactobacillus composition in gut microbiomes displays distinct associations with geographical location, age, BMI, and gender of the individuals. Lactobacillus-microbiome configurations in western countries may represent recent reconfiguration of a primordial intestinal ecotype, among which a specific configuration of lactobacilli is positively associated with not only age and BMI, but also with multiple diseases and disease marker taxa.

Results

Lactobacillus prevalence in the human gut and Lactobacillus-specific microbiome configurations

Lactobacilli present in human stool may be either autochthonous species being shed, or allochthonous strains transiently acquired from food or the physical environment. We reasoned that transient carriage would be less of a factor in a very large dataset derived by culture-independent methods. The curatedMetagenomicData repository provides such a resource, comprising more than 5,700 fecal shotgun metagenome datasets from 34 studies. Of these, 21 are disease-microbiome studies, with 17 containing paired control and diseased samples⁴⁶ (detailed in Supplementary Table S1). The remaining 13 study cohorts included only apparently healthy individuals, either from specific nationalities/ethnicities or age-groups. These datasets have all been collated and analyzed in a uniform manner which virtually eliminates bioinformatic-analysis-

related variations across studies. We supplemented this dataset with 408 fecal microbiome profiles from our own ELDERMET project and a recently published case-control dataset comprising of IBD patients.^{47,48}

We first explored the prevalence of various *Lactobacillus* species across the 6,155 collated fecal microbiome datasets. Overall, 2141 of the 6155 samples harbored at least one *Lactobacillus* species, detected with a relative abundance of 0.01% (See Methods for the selection of this threshold of detection). We next used a linear regression-based strategy that quantified the association of various host-associated demographic factors with the *Lactobacillus* detection rate (that is the number of *Lactobacillus* species detected per sample) after taking into account the study-specific (technical) variations (by especially taking the study name as a confounder) (Supplementary Table S2a). The detection rates were significantly associated with geography (country), age-group and the study-conditions of the individuals. Study condition refers to the clinical status of the individual from whom the corresponding gut microbiome sample was collected (as part of the original study and then collated in the curatedMetagenomicData repository). Study condition indicates whether an individual is an apparently healthy control, or is suffering from a specific disease or has undergone a specific treatment like antibiotics or fecal microbiome transplantation (FMT). Thus, the above result indicates that even after adjusting for study-specific factors, the overall prevalence rates of *Lactobacilli* showed significant variation, not only with respect to the country or the age-group of the individuals, but also with clinical status. Furthermore, using PERMANOVA analysis, we observed that these associations remain significant even at the level of the abundance of the individual species, after accounting for the study-specific technical factors like DNA extraction method, sequencing depth, and sequencing methodology (Supplementary Table S2b). Next, we focussed only on the subset of 4,303 non-diseased controls to investigate whether apparently healthy individuals were characterized by distinct configurations of gut *Lactobacilli* and whether (and how) these configurations varied with respect to the geography, age-group and other demographic factors.

Overall, 1,459 of the 4,303 (34%) of 'non-diseased' controls harbored at least one *Lactobacillus* species, (Supplementary Table S3). The detection pattern encompassed 47 *Lactobacillus* species, with 22 of these (hereafter referred to as 'rare' lactobacilli) detected in less than 5 of the 1,459 samples. We detected 25 *Lactobacillus* species above this threshold in 1,459 samples belonging to 31 cohorts from 22 countries (Figure 1(a)), with aggregate presence values ranging from 505 samples (*L. ruminis*), through 124 samples (*L. mucosae*), to 30 samples (*L. iners*) and below. For descriptive purposes, we divided the detected species into high (detected in greater than 100 samples), medium (detected in 50 to 100 samples) and low (less than 50 samples) prevalence groups (Figure 1(a)). The high-prevalence group included two species commonly consumed as probiotics, *L. casei* and *L. rhamnosus*, but the most prevalent species was *L. ruminis* which is also found in animals, and has the property of some strains being motile.⁴⁹ The species commonly used in combination with *Streptococcus thermophilus* for yogurt fermentation, *L. delbrueckii*, was the fourth most prevalent species (Figure 1(a)). The medium prevalence lactobacilli comprised three species found in fermented foods, *L. acidophilus* (also used as a probiotic), *L. sakei*, and *L. plantarum*. The low prevalence *Lactobacillus* group included three species commonly found in the vagina (*L. iners*, *L. vaginalis*, and *L. jensenii*), the others being primarily food or animal associated species (Figure 1(a)). We also observed that 1013 of the 1459 samples were characterized by the presence of a single *Lactobacillus* spp. (69%), with only 6% having three or more species (Figure 1(b)).

To further clarify the *Lactobacillus* configurations in the gut microbiomes of apparently healthy individuals, we asked if the individuals could be clustered in terms of their relatedness based on *Lactobacillus* species abundance. This approach is conceptually similar to enterotypes⁵⁰ but based on *Lactobacillus* abundance profiles. This primarily consisted of two steps, the first being the identification of an optimal number of clusters and the second being identification of the key species associated with each cluster. For identification of the optimal number (k) of clusters, we adopted an iterative approach, wherein we performed 100

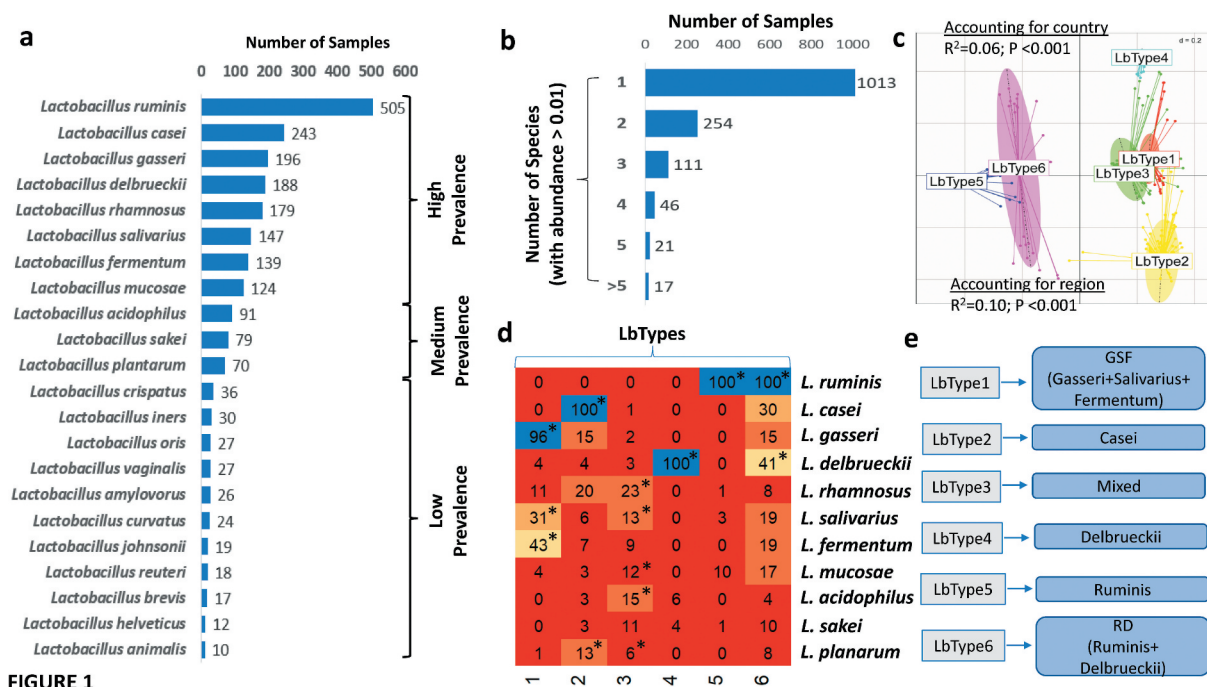


FIGURE 1

Figure 1. The gut microbiome of apparently healthy individuals is characterized by distinct lactobacillus populations. **a.** Bar plots showing the number of times each *Lactobacillus* species was detected (with abundance > 0.01%) in samples from ‘non-diseased’ control individuals. The species belonging to the ‘High-Prevalence’, ‘Medium-Prevalence’ and ‘Low-Prevalence’ groups of lactobacilli are demarcated. **b.** Bar showing the number of control individuals having different number of *Lactobacillus* species in their gut microbiomes. **c.** Principal Component Analysis (PCoA) plots (with the top 2 PCoA axes) showing the samples belonging to the different *Lactobacillotypes* (LbType). R-Squared and P-values of the PERMANOVA analysis testing the significance of these splits after considering both the country and the region as confounders is indicated. **d.** Heatmap on the left panel shows the percentage occurrence of each High and Medium prevalent *Lactobacillus* species in the six LbType. The percentage occurrence is the calculated from the number of times a given species was detected in a sample belonging to the given LbType divided by the total number of samples belonging to the LbType. Species that are significantly enriched in certain LbType were identified using Fishers’ Exact Test (estimate > 1 and p-value < 0.05) are indicated with *. **e.** Based on the significant enrichment patterns, the six LbTypes were named as described based on the dominant species in each cluster.

iterations. In each iteration, we randomly selected 50% of the samples and computed the silhouette scores for different cluster numbers (ranging from 2 to 20). The distribution of the silhouette scores across the 150 iterations for each cluster number is shown in Supplementary Figure S1a. The highest median silhouette scores were obtained when the clusters were $k = 5$ or 6 . However, for $k = 6$, the variations across iterations was noticeably lower, indicating that the microbiomes could be clustered optimally into six clusters that are stable to variations across iterations. This identified the optimal number of clusters as six (Supplementary Figure S1b). We labeled these microbiome clusters as ‘*Lactobacillotypes*’ (or ‘LbTypes’) (Figure 1(c)). We further tested the robustness of these *Lactobacillotype* groupings using PERMANOVA analysis and Random Forest models. Regional factors capture variations in ethnicity, diet, lifestyle,

and other socio-economic status, and have been shown in previous studies to have to have the strongest effect on the microbiome variations. ^{43,48} First, we observed that the variations in the compositional abundance of *Lactobacilli* across the different *Lactobacillotypes* remained significant after accounting for country of origin (R-squared = 0.06 and $P < .001$), continental region (R-squared = 0.10 and $P < .001$). The associations remained significant even after accounting study name as a confounder (PERMANOVA R-squared = 0.07 and $P < .001$), thus indicating that study-specific technical variations do not influence these associations).

We next validated the LbTypes using 100 iterations of Random Forest models, wherein for each iteration, we randomly selected 50% of the 1,462 samples to act as the training subsets to predict the *Lactobacillotypes* based on the *Lactobacillus* profiles

and subsequently tested this, trained on the remaining 50% of the samples. The models achieved an overall prediction accuracy of 98.4% across all LbTypes (Supplementary Figure S2a). The group-specific accuracies indicated similar levels of cluster granularities for almost all the LbTypes (accuracies of greater than 97%), except for LbType 6 (having the lowest group-specific accuracy at 93.5%).

Random Forest models provide the importance scores quantifying the predictive power of each feature (in this the case the lactobacillus abundance) in the LbType classification scheme. These species-specific feature importance scores indicated four species, *L. ruminis*, *L. gasseri*, *L. casei*, *L. delbrueckii*, as the top predictors of the Lactobacillotype classifications (Supplementary Figure S2b). These can be regarded as the signature taxa for each LbType. Three Lactobacillotypes (LbType 2, LbType 4 and LbType 5) were associated with each of the top predictor Lactobacillus species identified using the Random Forest models (with each signature species being present in at 100 of all samples in the corresponding LbTypes). To probe this further, we then identified the Lactobacilli significantly enriched in each of the Lactobacillotypes using Fishers' exact test (Figure 1(d)). This combination of Random Forest models and Fishers' exact tests enabled a clear identification of LbType-specific signatures. While LbType 1 was observed to be mixed, enriched with *L. gasseri* (present in 96% of the samples) (along with *L. salivarius* and *L. fermentum* present in 43% and 31% of the samples belonging to this LbType, respectively), LbTypes 2, 4 and 5 were each linked to *L. casei*, *L. delbrueckii*, and *L. ruminis*, respectively (each characteristic species being present in 100% of the samples belong to the corresponding LbType) (Figure 1(d)). Consequently, the LbTypes 1, 2, 4 and 5 were respectively named as Gasseri/Salivarius/Fermentum (or GSF), Casei, Delbrueckii, and Ruminis (Figure 1(e)). There were two other LbTypes (LbType 3 and LbType 6). LbType6 was enriched with *L. ruminis* (present in all the samples belonging to this LbType, similar to LbType 5) but additionally associated with *L. delbrueckii* (present in 41% of the samples) (Figure 1(d)). Because of this reason, it was named as Ruminis/Delbrueckii (or RD) (Figure

1(e)) In contrast to other LbTypes, we could not detect any signature species for LbType 3. This LbType was characterized by multiple species (none of which were detected in more than 25% of the samples). These included *L. rhamnosus* (present in 23% of the samples), followed by *L. salivarius*, *L. acidophilus*, *L. mucosae*, *L. sakei* (all present between 10 and 15% of the samples belonging to this LbType) as well as *L. fermentum* (which was present in 9% of this LbTypes' samples). Consequently, LbType 3 was named as 'Mixed' (Figure 1(e)). (Hereafter these notations are used to refer to the LbTypes). The two *L. ruminis*-associated LbTypes were least related to the other LbTypes (Figure 1(c)).

We next investigated the association of these LbTypes with the global enterotypes (which we defined in a similar manner as the LbTypes but considering the global composition profiles of all the core taxa) across the 1,462 gut microbiome samples. All non-diseased individuals could be divided into three optimal clusters referred to as Enterotypes (or EnTypes 1–3) (Supplementary Figure S3a–c) associated with Prevotella, Bacteroides, and Bifidobacterium abundance, respectively (Supplementary Figure S3d; Supplementary Table S4). Using a combination of Fishers' exact test and logistic regression models (See Methods and Supplementary Table S5), we observed associations of certain Enterotypes with different regions and age-groups. These reflected previously reported associations of particular microbial taxa with different regions and age-groups. The associations reflected the previous known associations of specific gut microbial members with age and geography.^{51,52} These included significant enrichment of the Bifidobacterium-associated Enterotype 1 in infants; the enrichment of the Prevotella-associated Enterotype 3 in the Other (Non-industrialized) geography as well as the opposite trend observed for the Bacteroides-linked Enterotype 2. (Supplementary Figure S3d).

Analyzing the Lb-Type associations with the Enterotypes revealed a positive association between the ruminis-associated LbTypes 5 and 6 with the Prevotella-associated Enterotype 3 (Benjamini Hochberg FDR of Fishers' Exact Test < 0.05; Supplementary Figure S3e). A similar link was also found between the LbTypes 1 (GSF), 2

(Casei) and 3 (Mixed), and the Bifidobacterium-associated Enterotype 1 (Benjamini Hochberg FDR of Fishers' Exact Test < 0.05). The Bacteroides-associated Enterotype 2 was present across all LbTypes. However, significant positive associations were observed with LbTypes 3 (Mixed), 4 (Delbrueckii) and 5 (Ruminis) (Supplementary Figure S3e).

Discrete patterns of Lactobacillus prevalence by region and country

As described above, geographic variation incorporates the key variations in ethnicity, diet, and lifestyles, which have previously been identified as strong covariates of microbiome composition.^{43,44,48,52} However, previous studies of lactobacillus abundance in the human gut did not overtly consider the effects of geographical region and country.

The prevalence rates of Lactobacilli across the different regions displayed significant variations even after adjusting for study-specific effects (Supplementary Figure S4), with the nationalities belonging to the other non-industrialized regions having significantly higher prevalence rates as compared to North America (significantly lower rates of Lactobacilli prevalence).

We determined the proportion of individuals whose microbiomes could be assigned to each LbType as a function of geographical region (Figure 2(a)) and the proportional representation with each region of the LbTypes (Figure 2(b)). The European subjects were characterized by a significant enrichment of LbTypes Delbrueckii, Casei, Mixed, and Gasser/Salivarius/Fermentum (GSF) (Fishers' Exact Test Benjamini-Hochberg FDR 2.14e-2, 3.4e-22, 5.63e-2 and 1.04e-2, respectively), and a significant lower prevalence of the two

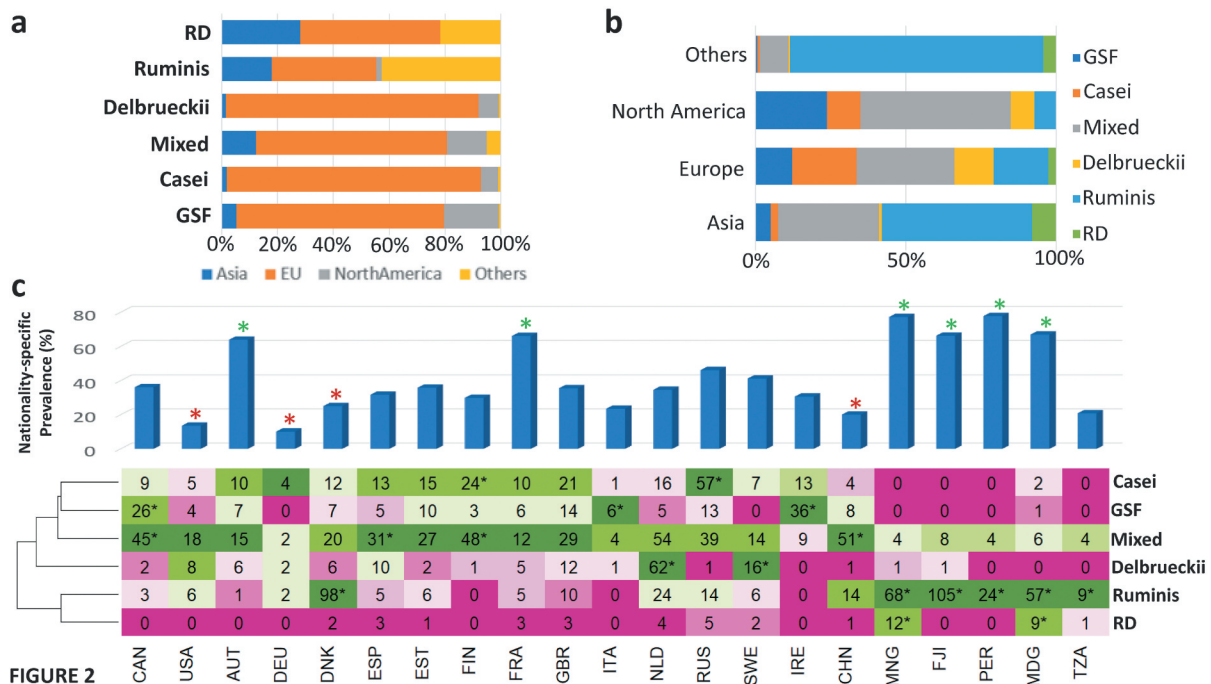


Figure 2. Lactobacillus species abundance in the gut displays significant region-specific trends. a. Stacked bar plots showing the region-wise percentage composition of each LbType b. Stacked bar plots showing the LbType percentage composition of each region. c. Top panel: Bar plots show the country-specific prevalence rates of Lactobacilli (percentage of samples from each country with at least one Lactobacillus species detected at abundance > 0.01%). Asterisks * in green indicate countries with significantly high prevalence rate, * in red indicate countries with significantly low prevalence rates (as compared to all others) (all with Benjamini-Hochberg FDR < 0.1). This was identified using Fishers' exact tests (See Methods). Bottom panel shows the heatmap showing the number of times each LbType was detected across each country. * indicates significant enrichment of a LbType in a given country using Fishers' exact test approach with Benjamini-Hochberg FDR < 0.1 (See Methods). For a given country, the colors of the cells are assigned based on the ranked detection of a LbType in that country (green for the highest detected LbType and red for the lowest detected LbType). * indicates significant enrichment of a LbType in a given country using Fishers' exact test approach with Benjamini-Hochberg FDR < 0.1. For a given country, the colors of the cells were assigned based on the ranked detection of a LbType in that country (green for the highest detected LbType and red for the lowest detected LbType).

ruminis-associated LbTypes (Fishers' Exact Test Benjamini-Hochberg FDR for Ruminis: $9.87e-6$ and RD: $5.6e-2$). In contrast, the *L. ruminis*-associated LbTypes were enriched in Asian subjects (Fishers' Exact Test Benjamini-Hochberg FDR for Ruminis $< 7.53e-7$ and RD: $3.12e-3$) and subjects belonging to the non-industrialized nationalities of Fiji, Peru, Tanzania and Madagascar (Fishers' Exact Test Benjamini-Hochberg FDR for Ruminis: $P < 5.11e-75$ and RD: 0.32).

European and North American subjects had more even distribution of a greater number of LbTypes and both harbored the Delbrueckii and Casei LbTypes which was noticeably absent in Asians and in the subjects from non-industrialized regions (Figure 2(a)). Over thirty percent of North American subjects were of the mixed LbType (Figure 2(b)). The North American subjects also displayed the highest prevalence of the Gasseri/Fermentum/Salivarius (GSF) LbType. Compared to other regions and subjects, the European subjects displayed the highest prevalence of the Delbrueckii and Casei LbType which may relate to yogurt and probiotic consumption, although paired dietary intake data for the study subjects was not available.

The highest prevalence rates by country (Figure 2(c)) were found in Austria, France, and in three non-Industrialized countries, Mongolia, Fiji and Peru. In France, the mixed and Casei LbTypes were the most prevalent. In Austria, the most prevalent LbType was Delbrueckii. In other European countries, there were clear and sometimes surprising differences; Danish subjects were dominated by the Ruminis LbType whereas in Finland and Estonia, Casei was more dominant. Spain showed a similar profile of Lactobacillotype abundance as France, but Great Britain differed due to higher relative abundance of the Gasseri/Salivarius/Fermentum Lactobacillotype. In the non-industrialized regions of Mongolia, Fiji, Tanzania, Madagascar, and Peru, the Ruminis LbType was significantly dominant. The clear influence of geography on the Lactobacillus composition of the gut microbiome was observed with respect to both the LbType to country associations (Chi-Square Test $P < 2.2e-16$) as well as the variation of the species profiles across different countries (after adjusting for Study-specific effects) (PERMANOVA analysis

of Species Composition Variance based on Kendall distances: $P < .001$).

These country-to-LbType relationships were also reflected in the country-to-species prevalence patterns (Supplementary Figure S5), whereby the Lactobacillus species composition of the subjects belonging to the North American and European countries were diverse (and characterized by distinctly lower prevalence of *L. ruminis*). In contrast, the subjects living in the non-industrialized countries were notably similar, being characterized by the dominance of *L. ruminis*.

Region-specific association of Lactobacilli with age, Body Mass Index (BMI) and gender

Given that geography was significantly associated with both the prevalence of *Lactobacillus* species and LbTypes, we performed a region-stratified analysis of the association of host anthropometric factors including age, BMI and gender on the Lactobacillus relative abundance in the gut microbiome.

Notwithstanding the fact that similarly-sized age groups were not equally distributed across all countries surveyed, in North America, the Lactobacillus prevalence rate in the Elderly being significantly higher than the Child/Teen/Young/Middle-aged groups (Figure 3(a); Fishers' exact $P < 2.2e-10$). Similarly, for the European individuals, the Elderly were observed to have significantly higher Lactobacillus prevalence as compared to the Child/Teen, Young and Middle-aged groups (Fishers' exact P -value < 0.002 ; Figure 3(a)). Only within Europe, the prevalence rates in the Infants were higher as compared to the Child/Teen/Young/Middle-aged groups (Fishers' exact P -value < 0.013). While infant samples were not available from the North American and Asian regions, in the other non-industrialized regions, no significance difference in Lactobacillus prevalence rates was observed between the infants and those belonging to the Child/Teen/Young/Middle-aged groups. In the non-industrialized regions, however, we observed significantly lower Lactobacillus prevalence rates in the elderly (as compared to the Child/Teen/Young/Middle-aged groups). This was in contrast to that observed

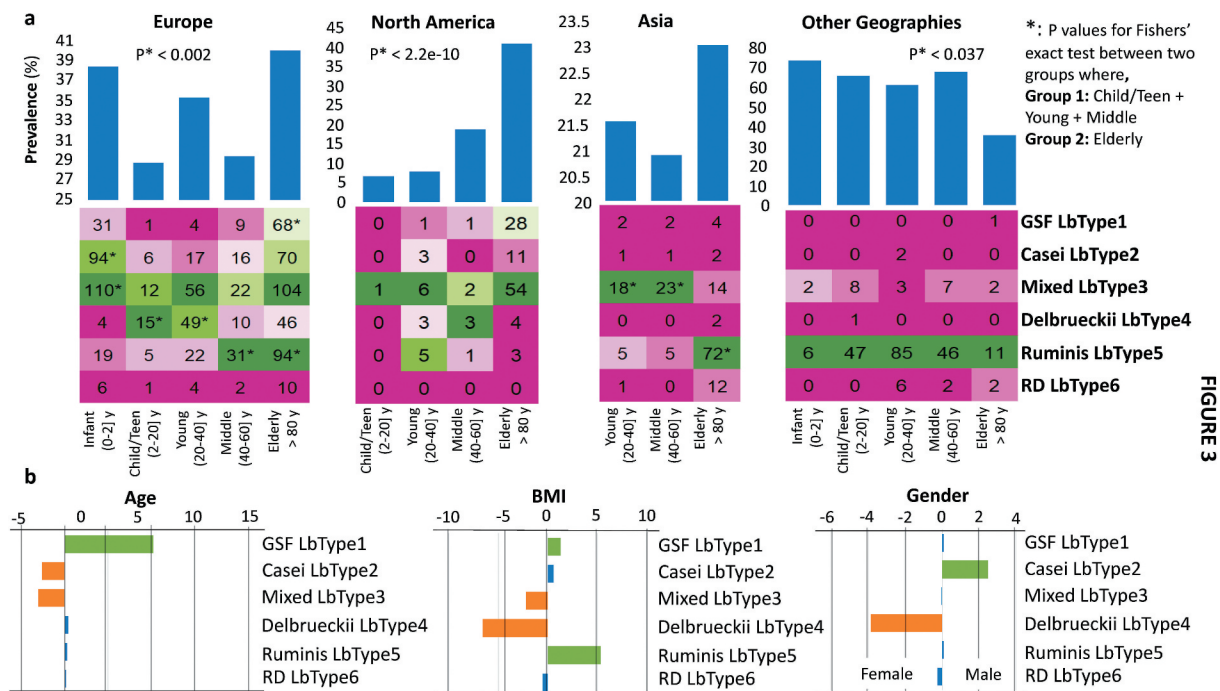


FIGURE 3

Figure 3. Host anthropometrics have specific associations with *Lactobacillus* species abundance, with age showing distinct region-dependent interaction. a. Bar plots on the top show the age-group specific prevalence rates of Lactobacilli across the four different regions. For a given region, the lactobacillus prevalence rates in a specified age-group were computed as the proportion of individuals (expressed as a percentage) in that age-group who have at least one lactobacillus species detected with abundance > 0.01 . Fisher's Exact Test p -values of the comparison of prevalence rates between Child/Young/Middle and the Elderly are shown for three regions of Europe, North America and the Other non-industrialized countries. Infants were not included in these groups because of the absence of infant samples from both North American and Asian regions, as well as the lack of consistent trends observed for the infant subjects across the other two regions. Heatmaps on the bottom panel show the number of times each LbType was detected in control individuals belonging to each age-group in each region. * indicates significant enrichment of a LbType in a given region. The age-groups of the individuals were defined as Infants: ≤ 2 years, Child/Teen: (2–20) years, Young: (20–40) years, Middle: (40–60) years and Elderly (> 60 years, with the maximum age being 102 years). b. Association extent computed using logistic regression (calculated by multiplying the $-\log_{10}$ of p -values (of base 10) with the directionality of the association) considering the region as a confounder. Positive associations (with $P < .05$) are indicated in green. Negative associations (with $P < .05$) are indicated in orange. See Methods for details on the regression analysis.

for the European and North American subjects. To further confirm that these observations were not consequences of country-specific biases in the proportional representation of age groups, we devised region-specific logistic regression models to compute the association of Lactobacillus prevalence rates with age after adjusting for country-specific variations (within each region) as confounders. The direction and the strength of the above associations were retained (however the negative association with age was not significant for the non-industrialized regions with $P < .1$) (Supplementary Table S7), confirming that Lactobacillus prevalence rates associate differentially with age depending upon the geographical location of the subject. Comparisons within the child/teen, young and

middle were neither significant nor yielded any coherent patterns across regions.

We also detected distinct region-specific enrichments for certain LbTypes in specific age-groups. In North America, this was reflected in higher prevalence of the Mixed, Gasseri/Fermentum/Salivarius and the Casei LbTypes in the elderly, while European elderly subjects gained these three LbTypes plus the Ruminis LbType (Figure 3(a), lower). For the Asian cohort, age information was only available for the Chinese individuals. In this cohort, the gain of lactobacilli with age was less clear-cut and was not statistically significant, perhaps because the cohorts available featured very low numbers of elderly subjects. However, we detected significant differences in the LbType composition across age-groups, with the Young and the

Middle age-groups being characterized by a significant enrichment of the mixed LbTypes, while the elderly were characterized by a significant gain of the Ruminis LbType. In the non-industrialized countries, an apparent decline in the Ruminis LbType prevalence was based on data for only 15 elderly subjects. Future analysis of more elderly subjects from these countries is required to investigate if our well-supported conclusion that Lactobacillus prevalence alters with age in North America and Europe also applies to those regions.

In addition to age, we also probed the association of Lactobacillus prevalence rates with BMI and gender separately within each geographical region. BMI data was available for 2092 subjects belonging to 16 studies (Supplementary Table S1). Increased lactobacillus abundance in obesity has been reported by one study,⁵³ but meta-analysis of the response of the gut microbiota to successful weight management interventions indicated that reduced lactobacillus abundance associated with weight loss.⁵⁴ We examined the prevalence of LbTypes as a function of Body Mass Index (BMI) in the gut metagenome data. For this we applied similar logistic regression models that checked the association of Lactobacilli prevalence first with BMI (and then with gender), while accounting for the country-specific variations as confounder (Supplementary Table S7). Region-stratified association analysis identified a marginally positive association between lactobacillus prevalence and BMI but only in European individuals (Logistic regression $R = 0.024$; $P < .054$) (Supplementary Table S7). This was not observed for any other geographical regions. However, there were variations in BMI ranges for the different studies with especially high BMI ranges for certain European cohorts like FengQ_2015 (median BMI = 27.6) and LeChatelierE_2013 (median BMI = 30.7) as compared to other regions. Such variations across region differences in BMI distributions may affect the ability to identify microbiome associations.

Further logistic regression analysis taking the region as cofounder identified *L. gasseri-salivarius-fermentum* LbType as positively associated with both age and BMI (Figure 3(b)). The mixed LbType showed the opposite trend. *L. delbrueckii* LbType was associated with lower BMI and the

female gender. Interestingly, the *L. ruminis* LbType was positively associated with BMI. However, these results should be treated with caution as they could not be validated after taking into account the country-specific biases within each region. We could not perform this adjustment because of the gender/BMI biases within countries and the complete absence of certain LbType in different countries.

Association between Lactobacillus abundance and disease

We next investigated the association of various Lactobacilli with different diseases. To avoid variations originating from differences in inclusion/exclusion criteria and experimental conditions, for each disease, we focussed only on cohorts specific to that disease. We identified 15 case-control studies corresponding to nine diseases that included at least 20 disease subjects and matched control samples (as part of the same study) (See Methods). We used logistic regression models to associate disease occurrence with the overall Lactobacillus prevalence rate as well as with the abundance of each *Lactobacillus* species after accounting for host anthropometric factors like age, BMI and gender.

We identified 19 significant Lactobacillus-Disease associations (with Benjamini-Hochberg $FDR < 0.1$) (Figure 4(a)). These encompassed 11 Lactobacillus species covering six out of the nine diseases. Out of these six, IBD, Cirrhosis, and type-II diabetes (T2D) were observed to have a significant increase in the overall detection of Lactobacilli, with IBD and Cirrhosis associated with increased prevalence rates of six Lactobacillus species, namely, *L. gasseri*, *L. salivarius*, *L. mucosae*, *L. delbrueckii*, *L. vaginalis*, and *L. oris*. T2D, on the other hand, was associated with increased prevalence of *L. amylovorus*. Polyps (adenoma) and colorectal cancer (CRC) were associated with a decrease in the Lactobacillus prevalence rates. Notably, Ruminis was associated with a decreased prevalence in multiple diseases, including cirrhosis, CRC, and Otitis. These associations could either indicate a direct link, or an indirect effect whereby altered lactobacillus abundance is associated with changes in the overall microbiome composition, be

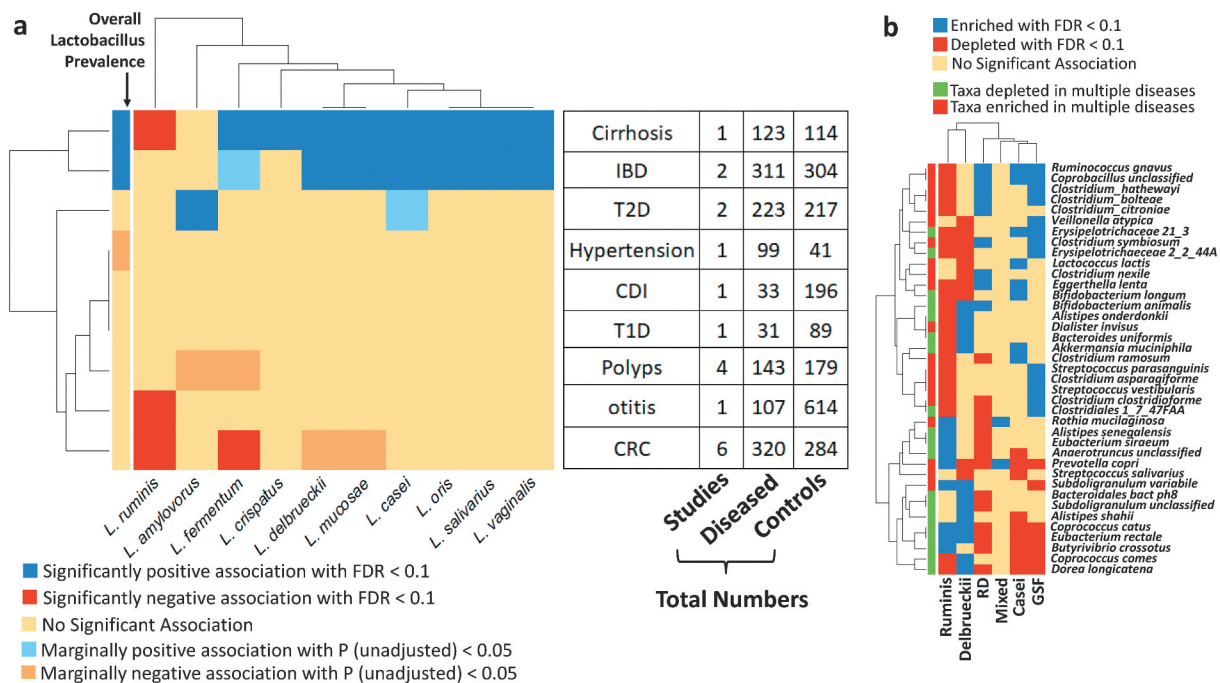


Figure 4. Lactobacillus species display distinct associations with multiple diseases and disease markers. a. Heatmap showing the significant association of the individual species as well as the overall *Lactobacillus* prevalence with the different diseases. The associations were obtained using Logistic regression models (within disease-specific country cohorts taking age, BMI and gender as confounders) (See Methods). Only those associations with either unadjusted $P < .05$ (shown as marginal associations) Benjamini-Hochberg FDR < 0.1 (as significant associations) are reported, along with the directionalities (as estimated from the regression models; that is positive or enriched if estimate > 0 and negative or depleted if estimate < 0). For each disease, we have shown the number of case-control studies included for each disease, the number of diseased subjects, and the number of control subjects. b. Heatmaps showing the association of the different generic disease markers with the different LbTypes identified using Mann-Whitney Tests (See Methods). Generic disease markers enriched or depleted in multiple diseases are indicated in green or red in the side color bars. Please refer to the original reference in Ghosh *et al.*⁴⁸ for details on the identification of these generic disease markers.

they causative or consequential of the indicated pathophysiology.

We next investigated, in the non-diseased control individuals, if certain Lactobacillotypes were more associated with disease-like microbiome configurations as compared to others. For this purpose, we checked if any of the gut microbial taxa previously shown to be associated with multiple diseases also displayed significant variations across the different Lactobacillotypes. In a recent meta-analysis covering five major diseases (namely, colorectal cancer, inflammatory bowel disease, type II diabetes, polyps and cirrhosis), we identified a specific group of taxa that were associated (either enriched or depleted) across multiple diseases (as shown in Figure 4(b)).⁴⁸ In addition, we also observed that a subset of these taxa that were enriched in multiple diseases were also associated with frailty in the ELDERMET cohort. In this current analysis, we, therefore, checked for either

enrichment or depletion of each of these multiple-disease-associated taxa in each of the six LbTypes. We found that subjects harboring the *L. gasseri-fermentum-salivarius* Lactobacillotype were enriched for several taxa including *C. citroniae*, *C. symbiosum*, *C. bolteae*, *C. asparagiforme*, *C. symbiosum*, *Clostridiales bacterium 1_47FAA*, that were not only enriched in multiple diseases but also associated with increased frailty in the elderly individuals (Figure 4(b)).⁴⁸ In contrast, this LbType was negatively associated with multiple health-associated taxa including *C. catus*, *E. rectale*, *B. crossotus*, *D. longicatena*. These data corroborate the association of *L. gasseri*, and *L. salivarius* with the altered microbiome found in multiple disease states (in other words as indicated by their increased detection in multiple diseases as seen in Figure 4(a)). In contrast, the *L. ruminis* Lactobacillotype showed the exact opposite trend, with enrichment of multiple health-associated taxa

and depletion of the pathobionts that were enriched in subjects of the Gasseri/Fermentum/Salivarius LbType. The Delbrueckii and Casei LbTypes also showed multiple positive and negative associations with health-associated taxa, respectively.

Discussion

Prior to the recent upsurge in culture-independent analyses of the gut microbiome, Lactobacilli were considered textbook examples of dominant or sub-dominant taxa. The vast number of 16S rRNA gene profiling studies and shotgun studies of the past decade showed this not to be true, but there had been no systematic studies of what Lactobacilli were present in what populations, and what, if any, host metadata co-varied with the Lactobacillus composition of the gut microbiome. Such insights could be helpful in the design and development of Lactobacillus-based (Lb-based) probiotic formulations. In spite of the availability of a multitude of over-the-counter Lb-based probiotic formulations, investigations into their clinical efficacy have yielded conflicting results, that are further confounded by the geographical region of the study-cohort, as well as by other host-associated factors like age (reviewed in⁵⁵). Studies in animals and humans have indicated that the ability of an administered probiotic to engraft is dependent on predictive baseline host and microbiome features.⁵⁶ It is debatable whether or not probiotic engraftment is desirable or necessary, but the specific patterns of lactobacillus species-gut microbiome interactions we report here are consistent with the notion that an administered probiotic will show different rates of successful network interaction with different microbiome types. The efficacy of any species to ameliorate disease symptoms may thus be modulated by the diet, environment and the indigenous microbiome of the host. Thus, insights into the differential prevalence of Lactobacilli across the geographical and age-landscape may aid in the formulation of personalized or population-specific probiotic formulations.

In this context, a key challenge is the accurate species-level identification of the Lactobacilli in the gut microbiomes. Given their close phylogenetic relatedness, species assignment of lactobacilli is challenging based on 16S rRNA amplicon data

(specifically those using short-read sequencing technologies) and so we relied on shotgun metagenomic sequencing data for the current study. The thresholds for the detection of each species has been validated for the Metaphlan software,⁵⁷ but our thresholds erred on the side of caution, so we acknowledge the possibility of under-detection of some species. Despite this technical caveat, the rank order of species prevalence aggregated across the subjects analyzed is broadly in line with older culture-based analyses (and reviewed in Introduction). We detected *L. buchneri* in only 9 samples above 0.01% abundance (Supplementary Table S1) although it was commonly cultured by Reuter,²¹ which probably reflects the fact that it is isolated from pressed yeast, milk, cheese, and fermenting plant material and this could vary dramatically by time period and geography. *L. reuteri* was also detected at relatively low prevalence compared to historical culture-based data, being found in the current study in only 18 subjects above 0.01% relative abundance. Both *L. reuteri* and *L. buchneri* were sparsely detected in the Non-industrialized populations (0 for *L. reuteri* and 2 for *L. buchneri*) (Supplementary Figure S5). This is in line with studies by Walter that suggest human *L. reuteri* strains are a bottle-necked clonal line due to a relatively recent dramatic reduction in carriage of this species,²² perhaps driven by industrialized Western diet.

The Enterotypes identified in the current study largely capture some of the known patterns of gut microbiome variation with age and region. The interesting insights obtained in the current study pertain to the specific LbType-Enterotype associations. *L. ruminis* is one of the few species suggested by Reuter to be autochthonous in humans which is consistent with our finding that it is by far the most highly prevalent species.²¹ As well as being present in all major geographical regions that we could survey, this species was particularly prevalent and abundant in non-industrialized countries, where it correlated with the relative over-abundance Prevotella-associated enterotype 3. The genus *Prevotella* displays important strain-level diversity in phenotype not investigated here,⁵⁸ but the overall abundance is driven by fiber and complex carbohydrate intake, whereas *L. ruminis* strains found in humans, as well as being non-motile compared

to animal isolates, have modest carbohydrate degradation capacity,⁵⁹ as do many lactobacillus species found in the human gut.⁶⁰ However, *L. ruminis* has been shown to effectively metabolize tetrasaccharides released from more complex substrates by gut microbiota taxa including *Coprococcus catus*.⁶¹ This was further confirmed in the current study wherein we observed a significant enrichment of *C. catus* along with other fiber-degrading bacteria like *E. rectale*, *B. crossotus* in individuals belonging to the *L. ruminis* LbType. Thus, the prevalence and abundance of *L. ruminis* in non-industrialized subjects could reflect its ability to cross-feed on substrates released from a high-fiber diet consumed by subjects in non-industrialized countries, who we recently showed have a very distinctive global microbiome compared to subjects in industrialized or Western countries.⁶²

The human gut microbiome changes with age, although age is a weaker co-variant with microbiome than region or country.⁴⁸ The age-dependent prevalence of lactobacillus is clearest in North America (increasing prevalence by age), which has the lowest fiber intake, whereas in the non-industrialized countries for which data was available the lactobacillus prevalence showed no noticeable difference across age into adulthood, but with a reduction in the elderly. This observation is largely based on the prevalence of the Ruminis and Ruminis/Delbrueckii (RD) LbTypes, and could also reflect diet, with declining fiber intake in the elderly. Similar trends of increasing prevalence with age are also reflected in Europe (from childhood till old age). However, a high prevalence in infancy could reflect probiotic consumption, especially of *L. casei* and *L. rhamnosus*. Previous studies investigating dietary differences have identified highest intake of dairy-based products in Europeans.⁶³

Our analyses of lactobacillus abundance interaction with disease does not allow distinction between cause and consequence. The enrichment of several lactobacillus species with cirrhosis, IBD, fatty liver disease and impaired glucose tolerance, viewed in isolation, present equivocal evidence for a role in disease. Based on our recent analysis of microbiome alterations in 2,500 case-control subjects, we studied lactobacillus associations with other genera and species that we showed displayed

age-specific abundance changes in multiple diseases.⁴⁸ This revealed that the lactobacillus species that are higher or lower in abundance in the diseases tested show strong associations with the broader microbiome changes that are characteristic of these diseases, especially with the abundance of the microbiome signature taxa for these diseases. Thus, it is likely that the lactobacillus-disease associations are reflective of the overall gut ecological changes in these diseases. However, at least one medical case report exists where in injection of multiple Lb species including *L. salivarius* and *L. fermentum* was linked with primary biliary cirrhosis (in line with the results obtained in the current study albeit at the gut microbiome level).⁶⁴ Furthermore, these hitherto unreported associations between LbTypes and the known disease markers indicate the need to consider the baseline state of an individual's microbiome in addition to other anthropometric factors prior to specific Lb-based probiotic administrations.

Different kinds of host-associated and technical factors are likely to affect the results of such meta-analyses that combine multiple datasets. We attempted to systematically account for a majority of these factors. First, we show that all the major host-associated factors (geography, age-group and disease) investigated in this study show significant associations with Lactobacillus detection rates even after adjusting for study-specific technical factors (the data for which was available for the curatedMetagenomicData repository). In the disease-association analyses, we further restricted our analyses to the study cohorts specific to each disease. This precludes biases in the results originating from varying inclusion/exclusion criteria as well as variations in experimental methodology. However, there were other factors which could not be accounted or investigated because of the non-availability of these metadata in the curatedMetagenomicData repository. These include experimental/technical factors (like storage or transport of samples) as well as host-associated life-style/demographic data like polypharmacy information (which has a key influence on the composition of gut microbiome), ethnicity information pertaining to specific individuals belonging to a nationality as well as other information regarding dietary habits. Addition of these information in

future versions of shotgun fecal microbiome data repositories could help shed light on the role of these factors on future microbiome-host association studies. Nevertheless, the information obtained herein provides direction for future studies that could focus on these factors in greater detail.

The practice of employing lactobacillus cultures for various beneficial functions and processes has been complicated by challenges in species identity, nomenclature and relatedness that will soon be simplified by the long-overdue taxonomic and phylogenetic overhaul of the 253 species.^{24,65} Aided by the genome sequences of type species, and the ambition to sequence to sequence a million human gut microbiomes (“Million Microbiome of Humans Project” (<https://en.mgitech.cn/news/114/>)), we will soon have even greater geographic and age coverage to adjust for confounders in determining lactobacillus-microbiome-host interactions. However, the present study already shows that human gut lactobacilli reflect and respond to the geographic and lifestyle differences in the human host, and identify the key species involved.

Materials and methods

Collation of fecal microbiome datasets

The curatedMetagenomicData is available as a R package, downloadable from Bioconductor. The curatedMetagenomicData package (as on September 2019) was downloaded, filtered and processed as discussed in a previous study.⁴⁸ This created a repository of 5746 gut (fecal) microbiome profiles. To this, we added a further 189 and 219 from the ELDERMET cohort and a previous IBD case-control dataset.^{47,48} The samples from the later two datasets were processed using the same approach as used for the samples in the curatedMetagenomicData repository, using Metaphlan2 and humann2.^{66,67} The details of the datasets included in this study along with their geographical locations, the distribution availability of the age and BMI of the subjects, as well as the total number and the number of ‘control’ (apparently healthy) and ‘non-control’ (diseased and other conditions) have been listed in Supplementary Table S1. The details of the subject-

specific metadata (including BMI, age, etc) for each study were already collected in the curatedMetagenomic repository from the individual studies (and used directly in this analysis). The collated fecal microbiome profiles corresponded to samples from 36 studies, spanning 25 nationalities across Europe, North America, Asia and Africa/South America/Oceania (grouped together as Others), distributed across five major age groups ranging from infancy to 102 years of age. The non-industrialized versus industrialized terminology was adopted from a recent study published recently by our group that used the same curatedMetagenomicData repository,⁶² where in designations of industrialized and non-industrialized status were based on the classifications of the different nationalities by the United Nations Industrial Development Organization (UNIDO).

Determination of Lactobacillus composition of the fecal microbiome

The first challenge here was to determine an appropriate abundance threshold to report a given *Lactobacillus* species as detected in a given metagenome. Using simulated metagenomes, previous studies on the validation of Metaphlan had observed that for abundance values greater than 0.01%, there was a linear relationship between the Metaphlan-calculated and the actual abundance values.⁵⁷ Therefore, we used this abundance threshold to identify a given *Lactobacillus* as being a present in each metagenome. Using this threshold, we identified the number of times a given *Lactobacillus* was detected across the different gut microbiome samples as well as the number of *Lactobacilli* detected in each sample. We specifically focussed on the non-diseased control individuals (tagged as ‘control’ in the study-condition metadata in the curatedMetagenomicData) for this purpose.

For obtaining a *Lactobacillus* specific species profile, we retained only those species belonging to the *Lactobacillus* genus and removed abundance values of less than 0.01%. For the overall microbiome profile, we restricted the species abundance profile to only those species that were present with abundance > 0.01% in 5% percent of the samples.

Identification of Lactobacillotypes and Enterotypes and validation of Lactobacillotypes using Random Forest

Unlike previous approaches on Enterotyping that have either used Jensen-Shannon Divergence (JSD) or Dirichlet Mixture Models (DMMs) on the global microbiome profiles,^{50,68} in this study, we adopted a two-step strategy. First, we used a dimensionality reduction technique Principal Coordinate Analysis (PCoA) on the gut microbiome profile. The objective behind implementing a dimensionality reduction was reduction of the 'noise' associated with individual samples. Often, a group of samples, that may have an otherwise similar composition with respect to the large majority of the constituent taxa, may have large variations amongst each other because of aberrantly high or low abundance of a few minority taxa that may have no contribution to the overall grouping schema. Utilizing a dimensionality-reduction technique and then representing each sample in terms of its top three axes (ordination axes explaining the largest variation) is expected to reduce this noise. Thus, in this study, we first performed a PCoA on the samples based on the intra-sample Spearman distances and subsequently represented the samples. We used the `dudi.pco` function (available in the `ade4` package) and the `cor` function, both available in the R programming interface.

The next step was to identify an optimal number of clusters that achieved the best grouping of the samples. In this case, there are indices like silhouette scores or the Calinski-Harabasz (CH) index that can measure the clustering efficiency of a grouping based on 'k' clusters. Computing these scores for different values of 'k' and identifying the 'k' achieving the best efficiency was the approach that have been used in earlier studies on Enterotyping.⁵⁰ However, performing the step of identifying the optimal 'k' just once on the entire set of samples may not be appropriate, as it can also be biased by the presence of outliers in this specific set. A way of addressing would be to repeatedly subsample the set (taking a subset) and repeat the step of identifying the optimal 'k' for each iteration. The final 'k' would then be identified as the one that not only achieves a high clustering efficiency, but the clustering efficiency to repeated iterations. We adopted this approach, where in, we performed 150

iterations, and in each iteration, we computed the silhouette scores for 'k' = 2 to 'k' = 20. Clustering was performed using k-means using the Euclidean distances of the top three PCoA axes. The value of 'k' that passed the two criteria of high as well as stable clustering efficiency across the 150 iterations was identified as the optimal number of Lactobacillotypes or Enterotypes. These Lactobacillotypes and Enterotypes were further performed and validated visually using heatmaps (`heatmap.2` function of the `gplots` package in R and the `cutree` function of the `dendextend` package) showing the clustering of the samples based on the Euclidean distances of the PCoA axes.

For the Lactobacillotypes, we performed an additional round of validation using Random Forest models wherein we again performed 100 iterations, where in each iteration, we trained models on 50% of the samples (for predicting the associated Lactobacillotype from the Lactobacilli composition) and tested the model on the rest 50%. The average classification accuracy per LbType as well as the full accuracy of classification was measured.

Associations between Enterotypes and the different regions and age-groups were performed using a combination of Fishers' exact test and logistic regression models. While Fishers' exact tests tested for association using simple count data, logistic regressions could test for the strength of these associations after taking into account the biases due to the various confounders (for example, biases in the representation of various age-groups across regions and biases in the representation of the different regions in sub-cohorts of individuals belonging to the different age-groups; described in detail in Supplementary Table S3).

Profiling regional variations of the Lactobacillus populations in gut microbiome

The association of specific Lactobacillotypes with distinct regions or countries were tested using Fishers' exact test. For each of these tests, a 2×2 contingency matrix was utilized which contained four values, namely the number of times a Lactobacillotype was detected in samples belonging to a given region/country, the remaining number of samples in the region/country, the number of

times a Lactobacillotype was detected in all the other regions/countries, the remaining number of samples (that is those in which the Lactobacillotype is not detected) in all the other regions/countries. Fishers' exact test provides two measures, namely the estimate (the extent of enrichment in the region/country versus the others, greater than 1 indicating enrichment and less than one indicating depletion) and the p -value (indicating the significance of association). For any region or country, enriched Lactobacillotypes were identified as those having estimates of greater than 1 and Benjamini–Hochberg FDR < 0.1.

For comparing the prevalence rates across countries, we used similar Fishers' exact test-based approach, wherein we counted the total number of samples where in any lactobacillus was detected rather than a specific Lactobacillotype. Specifically, a 2×2 contingency matrix was utilized which contained four values, namely the number of times any Lactobacillus was detected in samples belonging to a given region/country, the remaining number of samples in the region/country, the number of times any Lactobacillus was detected in all the other regions/countries, the remaining number of samples (that is those in which Lactobacilli were not detected) in all the other regions/countries. For any region or country, enriched prevalence of lactobacilli was identified when estimates of greater than 1 were obtained along with Benjamini–Hochberg FDR < 0.1.

Association analysis of Lactobacillotypes with host anthropometrics and disease

Association of *Lactobacillus* prevalence and Lactobacillotypes with host anthropometrics (age-group/age/BMI/gender) and disease were investigated using Fishers' exact tests (as described above) and validated using Logistic regression models (that took into account various host-associated factors like region, country, age, BMI, and gender). The age-groups of the individuals were defined as Infants: ≤ 2 years, Child/Teen: (2–20] years, Young: (20–40] years, Middle: (40–60] years and Elderly (> 60 years, with the maximum age being 102 years). The selection of these host-associated factors as

possible de-confounders was done based on the results of our previous meta-analysis,⁴⁸ wherein these host-associated metadata were present in more than 30% of the subjects, and also had the highest effect on the microbiome composition.

The region-stratified logistic regression models (separate models for each region) associating Lactobacillus prevalence with host anthropometrics were obtained as follows:

glm(LactobacillusDetected (1: Detected and 0: Not Detected) ~ Country + Age/BMI/Gender, family = "binomial")

These models considered the country-wise variations as a confounder.

The association of Lactobacillotypes to host anthropometrics were computed as:

glm(Lactobacillotype X (1: If sample belongs to LbType X and 0: If it does not belong to LbType X) ~ Region + Age/BMI/Gender, family = "binomial")

These models considered the region-wise variations as a confounder.

We next probed the association of the overall prevalence or the different Lactobacilli with disease. First, we removed those disease datasets from this analysis that contained fewer than 20 diseased subjects. This resulted in a list of 11 diseases, namely, adenoma (or polyps), atherosclerosis (AS), Clostridium difficile infection (CDI), cirrhosis, colorectal cancer (CRC), hypertension, inflammatory bowel disease (IBD), otitis, premature-born, type-I diabetes (T1D) and type-II diabetes (T2D). Out of these 11 diseases, for AS and premature born individuals, there were no matched controls sequenced as part of the same study (Supplementary Table S1), thereby restricting us to focus on the remaining nine diseases for this analysis. This included 15 matched case-control study datasets. For each disease, we constituted disease-specific bins by collating 'control' and 'case' samples from the 'case-control' study datasets corresponding to that disease.

Subsequently, only within the disease-specific bins, we utilized logistic regression models (as follows):

glm(LactobacillusDetected~Age+BMI+gender +DiseaseStatus, method = "binomial") (for overall prevalence)

glm(Lactobacillus Species 'X' Detected~Age+BMI+gender+DiseaseStatus, method = "binomial") (for overall prevalence) (for a specific *Lactobacilli X*)

The list of multiple disease markers were obtained from a previous study by our group.⁴⁸ The association of generic disease markers with *Lactobacillotypes* was performed using Mann-Whitney Tests. For each *Lactobacillotype*, the abundances of the disease markers were compared between all samples belonging to a *Lactobacillotype* and those not belonging to the *Lactobacillotype*. Species that were significantly enriched or depleted with Benjamini-Hochberg FDR < 0.1 were identified.

Disclosure of potential conflicts of interest

No potential conflicts of interest were disclosed.

Funding

This study was funded by Science Foundation Ireland in the form of a research centre award (APC/SFI/12/RC/2273_P2) to APC Microbiome Ireland.

ORCID

Tarini Shankar Ghosh  <http://orcid.org/0000-0001-9570-0365>

Paul W. O'Toole  <http://orcid.org/0000-0001-5377-0824>

References

- Ley RE, Hamady M, Lozupone C, Turnbaugh PJ, Ramey RR, Bircher JS, Schlegel ML, Tucker TA, Schrenzel MD, Knight R et al. Evolution of mammals and their gut microbes. *Science*. 2008 Jun 20;320(5883):1647–1651. doi:10.1126/science.1155725.
- Rosenberg E, Zilber-Rosenberg I. Microbes drive evolution of animals and plants: the hologenome concept. *MBio*. 2016 Mar 31;7(2):e01395. doi:10.1128/mBio.01395-15.
- Engel P, Moran NA. The gut microbiota of insects - diversity in structure and function. *FEMS Microbiol Rev*. 2013 Sep;37(5):699–735. doi:10.1111/1574-6976.12025.
- Hooks KB, O'Malley MA. Dysbiosis and its discontents. *MBio*. 2017 Oct 10;8:5. doi:10.1128/mBio.01492-17.
- Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. *Nat Rev Genet*. 2012 Mar 13;13(4):260–270. doi:10.1038/nrg3182.
- Flint HJ, Scott KP, Louis P, Duncan SH. The role of the gut microbiota in nutrition and health. *Nat Rev Gastroenterol Hepatol*. 2012 Oct;9(10):577–589. doi:10.1038/nrgastro.2012.156.
- Yang L, Lu X, Nossa CW, Francois F, Peek RM, Pei Z. Inflammation and intestinal metaplasia of the distal esophagus are associated with alterations in the microbiome. *Gastroenterology*. 2009 Aug;137(2):588–597. doi:10.1053/j.gastro.2009.04.046.
- Zackular JP, Baxter NT, Iverson KD, Sadler WD, Petrosino JF, Chen GY, Schloss PD, et al. The gut microbiome modulates colon tumorigenesis. *MBio*. 2013 Nov 5;4(6):e00692–13.
- Holmes E, Li JV, Athanasiou T, Ashrafiyan H, Nicholson JK. Understanding the role of gut microbiome-host metabolic signal disruption in health and disease. *Trends Microbiol*. 2011 Jul, 19(7):349–359.
- Mihaila D, Donegan J, Barns S, LaRocca D, Du Q, Zheng D, Vidal M, Neville C, Uhlig R, Middleton FA. The oral microbiome of early stage Parkinson's disease and its relationship with functional measures of motor and non-motor function. *PLoS One*. 2019;14(6):e0218252. doi:10.1371/journal.pone.0218252.
- Flemer B, Warren RD, Barrett MP, Cisek K, Das A, Jeffery IB, Hurley E, O'Riordain M, Shanahan F, O'Toole PW, et al. The oral microbiota in colorectal cancer is distinctive and predictive. *Gut*. 2018 Aug;67(8):1454–1463. doi:10.1136/gutjnl-2017-314814.
- Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, Almeida M, Arumugam M, Batto J-M, Kennedy S et al. Richness of human gut microbiome correlates with metabolic markers. *Nature*. 2013 Aug 29;500(7464):541–546. doi:10.1038/nature12506.
- Zeevi D, Korem T, Zmora N, Israeli D, Rothschild D, Weinberger A, Ben-Yacov O, Lador D, Avnit-Sagi T, Lotan-Pompan M et al. Personalized nutrition by prediction of glycemic responses. *Cell*. 2015 Nov 19;163(5):1079–1094. doi:10.1016/j.cell.2015.11.001.
- Hammes WP, Vogel RF. The genus *Lactobacillus*. In: Wood BJB, Holzapfel WH, editors. The genera of lactic acid bacteria. London (England): Blackie Academic & Professional. 1995;19–54.
- Stiles ME. Biopreservation by lactic acid bacteria. *Antonie Van Leeuwenhoek*. 1996 Oct;70(2–4):331–345. doi:10.1007/BF00395940.
- Abdel-Rahman MA, Sonomoto K. Opportunities to overcome the current limitations and challenges for efficient microbial production of optically pure lactic acid. *J Biotechnol*. 2016 Oct;20(236):176–192. doi:10.1016/j.jbiotec.2016.08.008.
- Lebeer S, Vanderleyden J, De Keersmaecker SC. Genes and molecules of *Lactobacilli* supporting probiotic action. *Microbiol Mol Biol Rev*. 2008 Dec;72(4):728–764. doi:10.1128/MMBR.00017-08.
- Sun Z, Harris HM, McCann A, Guo C, Argimón S, Zhang W, Yang X, Jeffery IB, Cooney JC, Kagawa TF,

- et al. Expanding the biotechnology potential of lactobacilli through comparative genomics of 213 strains and associated genera. *Nat Commun.* 2015;6(1):8322. doi:10.1038/ncomms9322.
19. Zheng J, Ruan L, Sun M, Gänzle M. A genomic view of lactobacilli and pediococci demonstrates that phylogeny matches ecology and physiology. *Appl Environ Microbiol.* 2015 Oct;81(20):7233–7243. doi:10.1128/AEM.02116-15.
 20. Duar RM, Lin XB, Zheng J, Martino ME, Grenier T, Pérez-Muñoz ME, Leulier F, Gänzle M, Walter J, et al. Lifestyles in transition: evolution and natural history of the genus *Lactobacillus*. *FEMS Microbiol Rev.* 2017 Aug 1;41(Supp_1):S27–S48.
 21. Reuter G. The *Lactobacillus* and *Bifidobacterium* microflora of the human intestine: composition and succession. *Curr Issues Intest Microbiol.* 2001 Sep;2(2):43–53.
 22. Walter J. Ecological role of lactobacilli in the gastrointestinal tract: implications for fundamental and biomedical research. *Appl Environ Microbiol.* 2008 Aug;74(16):4985–4996. doi:10.1128/AEM.00753-08.
 23. Dunne C, Murphy L, Flynn S, O'Mahony L, O'Halloran S, Feeney M, Morrissey D, Thornton G, Fitzgerald G, Daly C. Probiotics: from myth to reality. Demonstration of functionality in animal models of disease and in human clinical trials. *Antonie Van Leeuwenhoek.* 1999 Jul-Nov;76(1/4):279–292. doi:10.1023/A:1002065931997.
 24. Salvetti E, O'Toole PW. When regulation challenges innovation: the case of the genus *Lactobacillus*. *Trends Food Sci Tech.* 2107;66:187–194. doi:10.1016/j.tifs.2017.05.009.
 25. Savino F, Cordisco L, Tarasco V, Palumeri E, Calabrese R, Oggero R, Roos S, Matteuzzi D. *Lactobacillus reuteri* DSM 17938 in infantile colic: a randomized, double-blind, placebo-controlled trial. *Pediatrics.* 2010 Sep;126(3):e526–33. doi:10.1542/peds.2010-0433.
 26. Vitellio P, Celano G, Bonfrate L, Gobetti M, Portincasa P, De Angelis M. Effects of *bifidobacterium longum* and *Lactobacillus Rhamnosus* on gut microbiota in patients with lactose intolerance and persisting functional gastrointestinal symptoms: a randomised, double-blind, cross-over study. *Nutrients.* 2019 Apr 19;11(4):886. doi:10.3390/nu11040886.
 27. Zhao M, Shen C, Ma L. Treatment efficacy of probiotics on atopic dermatitis, zooming in on infants: a systematic review and meta-analysis. *Int J Dermatol.* 2018 Jun;57(6):635–641. doi:10.1111/ijd.13873.
 28. Kok CR, Hutkins R. Yogurt and other fermented foods as sources of health-promoting bacteria. *Nutr Rev.* 2018 Dec 1;76(Supplement_1):4–15. doi:10.1093/nutrit/nuy056.
 29. Marco ML, Heeney D, Binda S, Cifelli CJ, Cotter PD, Foligné B, Gänzle M, Kort R, Pasin G, Pihlanto A. Health benefits of fermented foods: microbiota and beyond. *Curr Opin Biotechnol.* 2017 Apr;44:94–102. doi:10.1016/j.copbio.2016.11.010.
 30. Arora T, Backhed F. The gut microbiota and metabolic disease: current understanding and future perspectives. *J Intern Med.* 2016 Oct;280(4):339–349. doi:10.1111/joim.12508.
 31. Million M, Angelakis E, Paul M, Armougom F, Leibovici L, Raoult D. Comparative meta-analysis of the effect of *Lactobacillus* species on weight gain in humans and animals. *Microb Pathog.* 2012 Aug;53(2):100–108. doi:10.1016/j.micpath.2012.05.007.
 32. Qin N, Yang F, Li A, Prifti E, Chen Y, Shao L, Guo J, Le Chatelier E, Yao J, Wu L. Alterations of the human gut microbiome in liver cirrhosis. *Nature.* 2014 Sep 4;513(7516):59–64. doi:10.1038/nature13568.
 33. Zegarra-Ruiz DF, El Beidaq A, Iniguez AJ, Lubrano Di Ricco M, Manfredo Vieira S, Ruff WE, Mubiru D, Fine RL, Sterpka J, Greiling TM et al. A diet-sensitive commensal *Lactobacillus* strain mediates TLR7-dependent systemic autoimmunity. *Cell Host Microbe.* 2019 Jan 9;25(1):113–127 e6. doi:10.1016/j.chom.2018.11.009.
 34. Halawa MR, El-Salam MA, Mostafa BM, Sallout SS. The gut microbiome, *Lactobacillus acidophilus*; relation with type 2 diabetes mellitus. *Curr Diabetes Rev.* 2019 Feb 6;15(6):480–485. doi:10.2174/1573399815666190206162143.
 35. Esmaeili SA, Mahmoudi M, Momtazi AA, Sahebkar A, Doulabi H, Rastin M. Tolerogenic probiotics: potential immunoregulators in systemic lupus erythematosus. *J Cell Physiol.* 2017 Aug;232(8):1994–2007.
 36. Dror T, Dickstein Y, Dubourg G, Paul M. Microbiota manipulation for weight change. *Microb Pathog.* 2017 May;106:146–161. doi:10.1016/j.micpath.2016.01.002.
 37. Bajaj JS, Heuman DM, Hylemon PB, Sanyal AJ, Puri P, Sterling RK, Luketic V, Stravitz RT, Siddiqui MS, Fuchs M. Randomised clinical trial: *Lactobacillus GG* modulates gut microbiome, metabolome and endotoxemia in patients with cirrhosis. *Aliment Pharmacol Ther.* 2014 May;39(10):1113–1125. doi:10.1111/apt.12695.
 38. Asgharian H, Homayouni-Rad A, Mirghafourvand M, Mohammad-Alizadeh-Charandabi S. Effect of probiotic yoghurt on plasma glucose in overweight and obese pregnant women: a randomized controlled clinical trial. *Eur J Nutr.* 2020 Feb;59(1):205–215. doi:10.1007/s00394-019-01900-1.019 May 8.
 39. Husni RN, Gordon SM, Washington JA, Longworth DL. *Lactobacillus* bacteremia and endocarditis: review of 45 cases. *Clin Infect Dis.* 1997 Nov;25(5):1048–1054. doi:10.1086/516109.
 40. Land MH, Rouster-Stevens K, Woods CR, Cannon ML, Cnota J, Shetty AK. *Lactobacillus* sepsis associated with probiotic therapy: fig 1. *Pediatrics.* 2005 Jan;115(1):178–181. doi:10.1542/peds.2004-2137.
 41. Yelin I, Flett KB, Merakou C, Mehrotra P, Stam J, Snesrud E, Hinkle M, Lesho E, McGann P, McAdam AJ. Genomic and epidemiological evidence

- of bacterial transmission from probiotic capsule to blood in ICU patients. *Nat Med.* 2019 Nov;25(11):1728–1732. doi:10.1038/s41591-019-0626-9.
42. He Y, Wu W, Zheng H-M, Li P, McDonald D, Sheng H-F, Chen M-X, Chen Z-H, Ji G-Y, Zheng ZDX, et al. Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nat Med.* 2018 Oct;24(10):1532–1535. doi:10.1038/s41591-018-0164-x.
 43. 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 Oct;24(10):1526–1531. doi:10.1038/s41591-018-0160-1.
 44. Claesson MJ, Jeffery IB, Conde S, Power SE, O'Connor EM, Cusack S, Harris HMB, Coakley M, Lakshminarayanan B, O'Sullivan O, et al. Gut microbiota composition correlates with diet and health in the elderly. *Nature.* 2012;488(7410):178–185. doi:10.1038/nature11319.
 45. Falony G, Joossens M, Vieira-Silva S, Wang J, Darzi Y, Faust K, Kurilshikov A, Bonder MJ, Valles-Colomer M, Vandeputte D et al. Population-level analysis of gut microbiome variation. *Science.* 2016 Apr 29;352(6285):560–564. doi:10.1126/science.aad3503.
 46. Pasolli E, Schiffer L, Manghi P, Renson A, Obenchain V, Truong DT, Beghini F, Malik F, Ramos M, Dowd JB et al. Accessible, curated metagenomic data through experimenthub. *Nat Methods.* 2017 Oct 31;14(11):1023–1024. doi:10.1038/nmeth.4468.
 47. Franzosa EA, Sirota-Madi A, Avila-Pacheco J, Fornelos N, Haiser HJ, Reinker S, Vatanen T, Hall AB, Mallick H, McIver LJ, et al. Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat Microbiol.* 2019 Feb;4(2):293–305. doi:10.1038/s41564-018-0306-4.
 48. Ghosh TS, Das M, Jeffery IB, O' Toole PW. Adjusting for age improves identification of gut microbiome alterations in multiple diseases. *eLife.* 2020. doi:10.7554/eLife.50240.
 49. Neville BA, Forde BM, Claesson MJ, Darby T, Coghlan A, Nally K, Ross RP, O'Toole PW. Characterization of pro-inflammatory flagellin proteins produced by *Lactobacillus ruminis* and related motile *Lactobacilli*. *PLoS One.* 2012;7(7):e40592. doi:10.1371/journal.pone.0040592.
 50. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto JM, et al. Enterotypes of the human gut microbiome. *Nature.* 2011 May 12;473(7346):174–180.
 51. Milani C, Duranti S, Bottacini F, Casey E, Turrone F, Mahony J, Belzer C, Delgado Palacio S, Arboleya Montes S, Mancabelli L, et al. The first microbial colonizers of the human gut: composition, activities, and health implications of the infant gut microbiota. *MMBR.* 2017 Dec;81:4.
 52. Sonnenburg ED, Sonnenburg JL. The ancestral and industrialized gut microbiota and implications for human health. *Nat Rev Microbiol.* 2019 Jun;17(6):383–390. doi:10.1038/s41579-019-0191-8.
 53. Drissi F, Raoult D, Merhej V. Metabolic role of lactobacilli in weight modification in humans and animals. *Microb Pathog.* 2017 May;106:182–194. doi:10.1016/j.micpath.2016.03.006.
 54. Seganfredo FB, Blume CA, Moehlecke M, Giongo A, Casagrande DS, Spolidoro JVN, Padoin AV, Schaen BD, Mottin CC. Weight-loss interventions and gut microbiota changes in overweight and obese patients: a systematic review. *Obes Rev.* 2017 Aug;18(8):832–851. doi:10.1111/obr.12541.
 55. Suez J, Zmora N, Segal E, Elinav E. The pros, cons, and many unknowns of probiotics. *Nat Med.* 2019 May;25(5):716–729. doi:10.1038/s41591-019-0439-x.
 56. Zmora N, Zilberman-Schapira G, Suez J, Mor U, Dori-Bachash M, Bashardes S, Kotler E, Zur M, Regev-Lehavi D, Brik RBZ. Personalized gut mucosal colonization resistance to empiric probiotics is associated with unique host and microbiome features. *Cell.* 2018 Sep 6;174(6):1388–1405.e21. doi:10.1016/j.cell.2018.08.041.
 57. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods.* 2012 Jun 10;9(8):811–814. doi:10.1038/nmeth.2066.
 58. De Filippis F, Pasolli E, Tett A, Tarallo S, Naccarati A, De Angelis M, Neviani E, Cocolin L, Gobetti M, Segata N. Distinct genetic and functional traits of human intestinal *Prevotella copri* strains are associated with different habitual diets. *Cell Host Microbe.* 2019 Mar 13;25(3):444–453.e3. doi:10.1016/j.chom.2019.01.004.
 59. Forde BM, Neville BA, O'Donnell MM, Riboulet-Bisson E, Claesson MJ, Coghlan A, Ross R, O'Toole PW. Genome sequences and comparative genomics of two *Lactobacillus ruminis* strains from the bovine and human intestinal tracts. *Microb Cell Fact.* 2011 Aug 30;10(Suppl 1):S13. doi:10.1186/1475-2859-10-S1-S13.
 60. McLaughlin HP, Motherway MO, Lakshminarayanan B, Stanton C, Paul Ross R, Brulc J, Menon R, O'Toole PW, van Sinderen D. Carbohydrate catabolic diversity of bifidobacteria and lactobacilli of human origin. *Int J Food Microbiol.* 2015 Jun;203(203):109–121. doi:10.1016/j.ijfoodmicro.2015.03.008.
 61. Lawley B, Sims IM, Tannock GW. Whole-transcriptome shotgun sequencing (RNA-seq) screen reveals upregulation of cellobiose and motility operons of *Lactobacillus ruminis* L5 during growth on tetrasaccharides derived from barley β -glucan. *Appl Environ Microbiol.* 2013 Sep;79(18):5661–5669. doi:10.1128/AEM.01887-13.

62. Keoghane 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 Jul 6;26(7):1089–1095. doi:10.1038/s41591-020-0963-8.
63. Singh GM, Micha R, Khatibzadeh S, Shi P, Lim S, Andrews KG, Engell RE, Ezzati M, Mozaffarian D. Global, regional, and national consumption of sugar-sweetened beverages, fruit juices, and milk: a systematic assessment of beverage intake in 187 countries. *PloS One.* 2015;10(8): e0124845. doi:10.1371/journal.pone.0124845.
64. Bogdanos D, Pusl T, Rust C, Vergani D, Beuers U. Primary biliary cirrhosis following *Lactobacillus* vaccination for recurrent vaginitis. *J Hepatol.* 2008 Sep;49(3):466–473. doi:10.1016/j.jhep.2008.05.022.
65. Zheng J, Wittouck S, Salvetti E, Franz CMAP, Harris HMB, Mattarelli P, O'Toole PW, Pot B, Vandamme P, Walter J, et al. 2020. A taxonomic note on the genus *Lactobacillus*: description of 23 novel genera, emended description of the genus *Lactobacillus* Beijerinck 1901, and union of *Lactobacillaceae* and *Leuconostocaceae*. *Int J Syst Evol Microbiol.* 2020 Apr;70(4):2782–2858. doi:10.1099/ijsem.0.004107.
66. Franzosa EA, McIver LJ, Rahnvard G, Thompson LR, Schirmer M, Weingart G, Lipson KS, Knight R, Caporaso JG, Segata N. Species-level functional profiling of metagenomes and metatranscriptomes. *Nat Methods.* 2018 Nov;15(11):962–968. doi:10.1038/s41592-018-0176-y.
67. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C, Segata N. MetaPhlan2 for enhanced metagenomic taxonomic profiling. *Nat Methods.* 2015 Oct;12(10):902–903. doi:10.1038/nmeth.3589.
68. Valles-Colomer M, Falony G, Darzi Y, Tigchelaar EF, Wang J, Tito RY, Schiweck C, Kurilshikov A, Joossens M, Wijmenga C. The neuroactive potential of the human gut microbiota in quality of life and depression. *Nat Microbiol.* 2019;Apr(4):623–632. doi:10.1038/s41564-018-0337-x.