Gut Microbiome Composition Is Predictive of Incident Type 2 Diabetes in a Population Cohort of 5,572 Finnish Adults

Diabetes Care 2022;45:811–818 | https://doi.org/10.2337/dc21-2358

Matti O. Ruuskanen,[1] Pande P. Erawijantari,[1] Aki S. Havulinna,[2,3] Yang Liu,[4,5] Guillaume Méric,[4,6] Jaakko Tuomilehto,[2,7,8] Michael Inouye,[4,9] Pekka Jousilahti,[2] Veikko Salomaa,[2] Mohit Jain,[10,11] Rob Knight,[12–15] Leo Lahti,[1] and Teemu J. Niiranen[2,16,17]

[1]Department of Computing, University of Turku, Turku, Finland
[2]Department of Public Health and Welfare, Finnish Institute for Health and Welfare, Helsinki, Finland
[3]Institute for Molecular Medicine Finland, Helsinki Institute of Life Science, Helsinki, Finland
[4]Cambridge Baker Systems Genomics Initiative, Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia
[5]Department of Clinical Pathology, Melbourne Medical School, University of Melbourne, Melbourne, Victoria, Australia
[6]Department of Infectious Diseases, Central Clinical School, Monash University, Melbourne, Victoria, Australia
[7]Department of Public Health, University of Helsinki, Helsinki, Finland
[8]Saudi Diabetes Research Group, King Abdulaziz University, Jeddah, Saudi Arabia
[9]Department of Public Health and Primary Care, Cambridge University, Cambridge, U.K.
[10]Department of Medicine, University of California San Diego, La Jolla, CA
[11]Department of Pharmacology, University of California San Diego, La Jolla, CA
[12]Jacobs School of Engineering, University of California San Diego, La Jolla, CA
[13]Center for Microbiome Innovation, University of California San Diego, La Jolla, CA
[14]Department of Pediatrics, School of Medicine, University of California San Diego, La Jolla, CA
[15]Department of Computer Science & Engineering, University of California San Diego, La Jolla, CA
[16]Division of Medicine, Turku University Hospital, Turku, Finland
[17]Department of Internal Medicine, University of Turku, Turku, Finland

Corresponding author: Matti O. Ruuskanen, matti.ruuskanen@utu.fi

Received 12 November 2021 and accepted 5 January 2022

This article contains supplementary material online at https://doi.org/10.2337/figshare.18092744.

M.O.R. and P.P.E. contributed equally to this work. L.L. and T.J.N. contributed equally to this work.

OBJECTIVE

To examine the previously unknown long-term association between gut microbiome composition and incident type 2 diabetes in a representative population cohort.

RESEARCH DESIGN AND METHODS

We collected fecal samples from 5,572 Finns (mean age 48.7 years; 54.1% women) in 2002 who were followed up for incident type 2 diabetes until 31 December 2017. The samples were sequenced using shotgun metagenomics. We examined associations between gut microbiome composition and incident diabetes using multivariable-adjusted Cox regression models. We first used the eastern Finland subpopulation to obtain initial findings and validated these in the western Finland subpopulation.

RESULTS

Altogether, 432 cases of incident diabetes occurred over the median follow-up of 15.8 years. We detected four species and two clusters consistently associated with incident diabetes in the validation models. These four species were *Clostridium citroniae* (hazard ratio [HR] 1.21; 95% CI 1.04–1.42), *C. bolteae* (HR 1.20; 95% CI 1.04–1.39), *Tyzzerella nexilis* (HR 1.17; 95% CI 1.01–1.36), and *Ruminococcus gnavus* (HR 1.17; 95% CI 1.01–1.36). The positively associated clusters, cluster 1 (HR 1.18; 95% CI 1.02–1.38) and cluster 5 (HR 1.18; 95% CI 1.02–1.36), mostly consisted of these same species.

CONCLUSIONS

We observed robust species-level taxonomic features predictive of incident type 2 diabetes over long-term follow-up. These findings build on and extend previous mainly cross-sectional evidence and further support links between dietary habits, metabolic diseases, and type 2 diabetes that are modulated by the gut microbiome. The gut microbiome can potentially be used to improve disease prediction and uncover novel therapeutic targets for diabetes.

The roles of host genetics and environmental factors in the pathogenesis of type 2 diabetes have been widely studied (1,2). Recently, several studies have reported a link between gut microbiome composition and type 2 diabetes (3–5). This association may involve several mechanisms, such as modulation of inflammation,

increased gut permeability, interactions with dietary constituents, glucose and lipid metabolisms, insulin sensitivity, and effects on overall energy homeostasis of the host (5). Specifically, type 2 diabetes has been reported to be associated with lower relative abundances of butyrate-producing microbes and increases in various opportunistic pathogens (4,6).

Most prior studies on the association between gut microbiome and type 2 diabetes have been limited by their cross-sectional designs (3,5). While these studies have begun to elucidate the role of the gut microbiome in type 2 diabetes pathogenesis, they are subject to selection bias and have not included prospective data on incident diabetes. As a result, such analyses provide limited information on how the gut microbiome could be used in the prediction of the development of diabetes. Prospective studies have thus far been conducted rarely, with short-term follow-up (7), or only in the context of diurnal oscillation of gut bacteria (8). In addition, growing evidence indicates that some previous results from cross-sectional studies might have been confounded by the use of antidiabetic drugs that can influence gut microbiome composition, such as metformin (9,10).

We analyzed the long-term association between gut microbiome composition and incident type 2 diabetes in a well-phenotyped and representative Finnish population sample (N = 5,572). The follow-up spanned 16 years after sampling (11). Notably, participants with prevalent diabetes at baseline, including those taking antidiabetic drugs such as metformin, were excluded from our study. The FINRISK 2002 cohort features participants both from eastern and western Finland with differences in genetics, lifestyle, and morbidity and mortality rates (12). To improve the robustness of our results, we performed feature selection separately in data from eastern Finland and evaluated the findings in participants from western Finland to establish robust microbial signals predictive of incident type 2 diabetes.

RESEARCH DESIGN AND METHODS

The FINRISK study has been conducted in Finland to investigate risk factors for cardiovascular disease every 5 years since 1972 (11). In 2002, the study included participants from six areas: North Karelia, Northern Savo, Oulu, Lapland, Turku and Loimaa, and Helsinki and Vantaa. These areas can be geographically divided roughly into western Finland (Turku and Loimaa and Helsinki and Vantaa), and eastern Finland categories (North Karelia, Northern Savo, Oulu, and Lapland). A random sample stratified by sex and 10-year age-groups among the population aged 24–74 years was taken in each study area. Of the 13,498 invitees, 8,783 participated in the study. Of these participants, 7,231 donated fecal samples. In the current study, we excluded individuals with one or more exclusion factors: prevalent diabetes (n = 698), pregnancy (n = 40), $<$50,000 mapped reads (n = 20), or antibiotic use in the past 6 months (n = 907). After these exclusions, samples from 5,572 participants were eligible for this study.

The health status of the participants was assessed at baseline in 2002 (11). Physical examination and blood sampling were performed at local health centers or other survey sites by nurses specially trained for the survey methods. Data were collected for physiologic measures, biomarkers, and dietary, demographic, and lifestyle factors (11). Willing participants were given a stool sampling kit with detailed instructions. Samples were mailed overnight under Finnish winter conditions to the laboratory, where they were immediately stored at -20°C. The samples were stored unthawed until 2017, when they were shipped to the University of California San Diego for sequencing. The Coordinating Ethics Committee of the Helsinki University Hospital District (Helsinki, Finland) approved the study protocol for FINRISK 2002 (ref. no. 558/E3/2001), and all participants provided written informed consent.

National health care registers in Finland enable combining of the data in FINRISK with subsequent in- and outpatient disease diagnoses and drug prescriptions based on individual personal identity codes. Prevalent and incident diabetes were defined based on ICD-10 codes E10–E14, ICD-9 code 250, or ICD-8 code 250 in the nationwide Care Register for Health Care. In addition, prevalent diabetes was based on three or more drug purchases with Anatomical Therapeutic Chemical drug code A10 in the nationwide Drug Reimbursement Register prior to baseline. This drug code (and therefore the exclusion) includes metformin, which is widely reported to alter gut microbiota (10). The register data were amended with the patient's self-report, measured fasting plasma glucose $\geq$7.0 mmol/L, 2-h oral glucose tolerance test plasma glucose $\geq$11.1 mmol/L, or HbA_{1c} $\geq$48 mmol/mol at baseline examination. A glucose tolerance test was available only for 3,378 participants and HbA_{1c} for 4,096 participants (of 5,572). The participants were followed through 31 December 2017.

Earth Microbiome Project protocols were used for DNA extraction with the MagAttract PowerSoil DNA Kit (Qiagen), as described previously (13). Library generation was performed with a miniaturized version of the Kapa HyperPlus Illumina-compatible library prep kit (Kapa Biosystems) (14). An Echo 550 acoustic liquid-handling robot (Labcyte, Inc.) was used to normalize DNA extracts to 5 ng total input per sample. With a Mosquito HV liquid-handling robot (TTP Labtech, Ltd), 1/10 scale enzymatic fragmentation, end-repair, and adapter-ligation reactions were performed. Sequencing adapters were based on the iTru protocol (15), where ligation of short universal adapter stubs is followed by addition of sample-specific barcoded sequences in a subsequent PCR step. PicoGreen assay was used to quantify amplified and barcoded libraries, which were pooled in approximately equimolar ratios before being sequenced on an Illumina HiSEq 4000 instrument. An average read count of 900,000 reads per sample was achieved with this protocol. Atropos was used for quality trimming of the sequences and removal of sequencing adapters (16). Bowtie2 (17) was used to remove host reads by mapping them against the human genome assembly GRCh38. SHOGUN version 1.0.5 (18) was used to assign taxonomy to the reads using National Center for Biotechnology Information RefSeq version 82 (8 May 2017), which contains complete bacterial, archaeal, and viral genomes, together with plasmid sequences.

All statistical analyses were performed with R version 3.6.1 (19). The data were first divided into participants from eastern Finland (n = 3,871) and western Finland (n = 1,701). These subpopulations were selected because of their well-known differences in genetic background, lifestyle, and mortality rate (12). Because of the larger number of

participants in eastern Finland, we used this data set to discover associations, followed by validation of the findings with western Finland data. α-Diversity of the microbiomes was assessed with raw counts per taxon and Shannon diversity. β-Diversity was calculated separately in the data from eastern and western Finland subpopulations by applying a centered log-ratio (CLR) transformation on the taxon counts followed by principal component (PC) analysis. Rare taxa were filtered out in eastern Finland data, with the cutoffs set at detection >0.1% and prevalence >10% of raw (untransformed) mapped reads. Taxa were then subset to this filtered set (119 taxa) and CLR transformed in all of the data.

Cox proportional hazards regression models for survival time were first constructed solely in data from eastern Finland with the R package survival version 3.2.11 (20). Models were constructed for *1*) observed counts (total number of raw taxon matches), *2*) Shannon diversity, *3*) first 10 PC axes (10 separate models), and *4*) relative taxon abundances (119 separate models). Each model was adjusted for baseline age, BMI, sex, systolic blood pressure, non-HDL cholesterol, triglycerides, and current smoking status of the participants. Features significantly associated with incident type 2 diabetes were filtered at α level $P < 0.05$ after applying Benjamini-Hochberg correction. Instead of correlation, we analyzed the compositionally valid measure, proportionality (ρ), between the significantly associated taxa using the R package propr version 4.2.6 (21). The taxa were then clustered based on proportionality with the Ward minimum variance method, and the optimal number of clusters was defined with Kelley-Gardner-Sutcliffe penalty function in the R package maptree version 1.4.7 (22). Heatmaps of the proportionality between taxa and associated clusters and hazard ratios (HRs) were visualized with the R package ComplexHeatmap version 2.7.11 (23). Relative abundances of the clusters were calculated by combining and CLR transforming the raw counts of the taxa within the data from eastern Finland.

Following the screening and selection of significant features in the data from eastern Finland, models were constructed identically and separately for western Finland. A feature selected in data from eastern Finland was considered to be robustly predictive of incident type 2 diabetes in the western subpopulation if the 95% CI of its HR did not overlap 1.0 (unadjusted $P < 0.05$). Finally, Kaplan-Meier curves were constructed for relative abundance quantiles of these robustly predictive features in data from western Finland with the R package rms version 6.2.0 (24).

Gut microbiomes of participants with undiagnosed type 2 diabetes at the baseline examination and sampling might have been affected by undiagnosed dysglycemia (9). It is, however, likely that these participants would have been diagnosed with type 2 diabetes during the early follow-up period. Therefore, an additional analysis was conducted by excluding participants diagnosed with type 2 diabetes within the first 2 years of follow-up.

Data and Resource Availability

The source code used to analyze the data and produce our results is included in the Supplementary Material and available under a permanent DOI in Zenodo: https://doi.org/10.5281/zenodo.5901114. Because of the sensitive health information of individuals, the data sets analyzed during the current study are not public but are available based on a written application to the Finnish Institute for Health and Welfare Biobank as instructed in https://thl.fi/en/web/thl-biobank/for-researchers.

RESULTS

The characteristics of the study participants are reported in Table 1. A total of 432 (7.8%) participants were diagnosed with type 2 diabetes over a median follow-up of 15.8 years.

In the data from eastern Finland, of the 119 taxa remaining after filtering, the relative abundances of 18 were significantly associated with incident type 2 diabetes (adjusted $P < 0.05$) (Figs. 1 and 2 and Supplementary Table 1). Fifteen taxa had positive associations with incident type 2 diabetes, and three taxa were negatively associated. Most of the positively associated taxa were from the family *Lachnospiraceae*, with several representatives of genus *Clostridium*. Two of the three negatively associated taxa were from genus *Alistipes*. α-Diversity was not significantly associated with incident type 2 diabetes (adjusted $P > 0.05$). In the β-diversity analysis, the first PC axis had a significant association (HR 0.82; 95% CI 0.69–0.88; adjusted $P = 0.01$). Significantly associated taxa could be grouped by proportional abundance into five clusters (Fig. 1). Four taxa and two clusters were positively associated with incident type 2 diabetes in the western Finland subpopulation (Fig. 2 and Supplementary Table 1). These taxa were *Clostridium citroniae* (eastern Finland: HR 1.21; 95% CI 1.09–1.35; western Finland: HR 1.21; 95% CI 1.04–1.42; unadjusted $P = 0.02$), *C. bolteae* (eastern Finland: HR 1.18; 95% CI 1.07–1.30; western Finland: HR 1.20; 95% CI 1.04–1.39; unadjusted $P = 0.01$), *Tyzzerella nexilis* (eastern Finland: HR 1.16; 95% CI 1.05–1.29; western Finland: HR 1.17; 95% CI 1.01–1.36; unadjusted $P = 0.03$), and *Ruminococcus gnavus* (eastern Finland: HR 1.18; 95% CI 1.06–1.30; western Finland: HR 1.17; 95% CI 1.01–1.36; unadjusted $P = 0.04$). The directions of these associations were the same as those in eastern Finland. Clustering the 18 selected taxa by proportional abundance separately in data of each subpopulation also produced clusters with identical taxon membership (Fig. 1). Three of the associated taxa, *C. citroniae*, *C. boltae*, and *R. gnavus*, were grouped in cluster 1 (western Finland: HR 1.18; 95% CI 1.02–1.38; unadjusted $P = 0.03$) with one additional taxon in the cluster, *E. lenta*, which was not associated with type 2 diabetes in western Finland data as an individual predictor. *T. nexilis* was grouped in cluster 5 (western Finland: HR 1.18; 95% CI 1.02–1.36; unadjusted $P = 0.03$) with two additional taxa in the cluster, *C. symbiosum* and *C. glycyrrhizinilyticum*, which were not individually associated with type 2 diabetes in western Finland data. In the β-diversity analysis, the first PC axis did not show an association with incident type 2 diabetes in the western Finland data (HR 0.94; 95% CI 0.79–1.11; unadjusted $P = 0.45$). Fewer participants in western Finland with a relative abundance of *C. citroniae* below quartile 1 (Q1) developed incident type 2 diabetes during the follow-up period than those above this quartile (Fig. 3). For all other microbial features, fewer participants in western Finland with a relative abundance below the median (Q2) of each feature developed incident type 2 diabetes than those with an abundance above the median.

In an additional analysis of the data, we excluded 44 participants (33 from

Table 1—Baseline statistics of the participants in FINRISK 2002 after exclusions

Variable	Total	Incident type 2 diabetes			Geographic area		
		Yes	No	*P**	Eastern Finland	Western Finland	*P**
Participants	5,572	432 (7.8)	5,140 (92.3)	—	3,871 (69.5)	1,701 (30.5)	—
Women	3,013 (54.1)	218 (50.5)	2,795 (54.4)	0.12	2,074 (53.6)	939 (55.2)	0.72
From eastern Finland	3,871 (69.5)	293 (67.8)	3,578 (69.6)	0.45	—	—	—
With incident type 2 diabetes	432 (7.8)	—	—	—	293 (7.6)	139 (8.2)	0.45
Baseline age, years	48.7 ± 12.8	52.8 ± 10.6	48.4 ± 13.0	1.0×10^{-11}	48.7 ± 12.9	48.7 ± 12.8	0.99
BMI, kg/m^2	26.6 ± 4.4	30.7 ± 5.2	26.3 ± 4.2	1.2×10^{-71}	26.8 ± 4.4	26.2 ± 4.4	3.4×10^{-7}
Systolic blood pressure, mmHg	135.2 ± 19.8	141.7 ± 20.4	134.6 ± 19.6	6.4×10^{-13}	135.9 ± 20	133.5 ± 19.2	2.8×10^{-5}
Non-HDL cholesterol, mmol/L	4.1 ± 1.1	4.5 ± 1.3	4.0 ± 1.1	1.1×10^{-14}	4.1 ± 1.1	4.0 ± 1.1	1.2×10^{-7}
0-h plasma glucose, mmol/L	5.7 ± 0.5	6.1 ± 0.5	5.7 ± 0.5	9.2×10^{-34}	5.7 ± 0.5	5.7 ± 0.5	1.1×10^{-3}
2-h plasma glucose, mmol/L	6.3 ± 1.7	7.5 ± 1.9	6.2 ± 1.6	8.3×10^{-24}	6.3 ± 1.7	6.4 ± 1.7	0.03
Hemoglobin A_{1c}, mmol/mol	35.8 ± 3.6	38.4 ± 3.8	35.5 ± 3.5	4.9×10^{-29}	35.6 ± 3.8	36.2 ± 3.1	7.9×10^{-5}
Triglycerides, mmol/L	1.4 ± 0.9	1.9 ± 1.3	1.3 ± 0.8	3.4×10^{-38}	1.4 ± 0.9	1.4 ± 0.9	0.11
Current smoking	1,327 (23.8)	111 (25.7)	1,216 (23.7)	0.38	914 (23.6)	413 (24.3)	0.63

Data are presented as *n* (%) (*n* of participants in indicated category and percentage of total) or mean ± SD. *Mann-Whitney *U* test was used for numeric data; Fisher exact test was used for categorical data.

western and 11 from eastern Finland) diagnosed with type 2 diabetes within the initial 2 years of follow-up. Of the 18 taxa passing the *P* value filtering in the full data from eastern Finland, 17 passed the same filter (adjusted $P < 0.05$) in the subset data (Supplementary Fig. 1 and Supplementary Table 2). Additionally, four other species passed this filter in the subset data in eastern Finland. These 21 species clustered by proportional abundance in the eastern Finland data into six

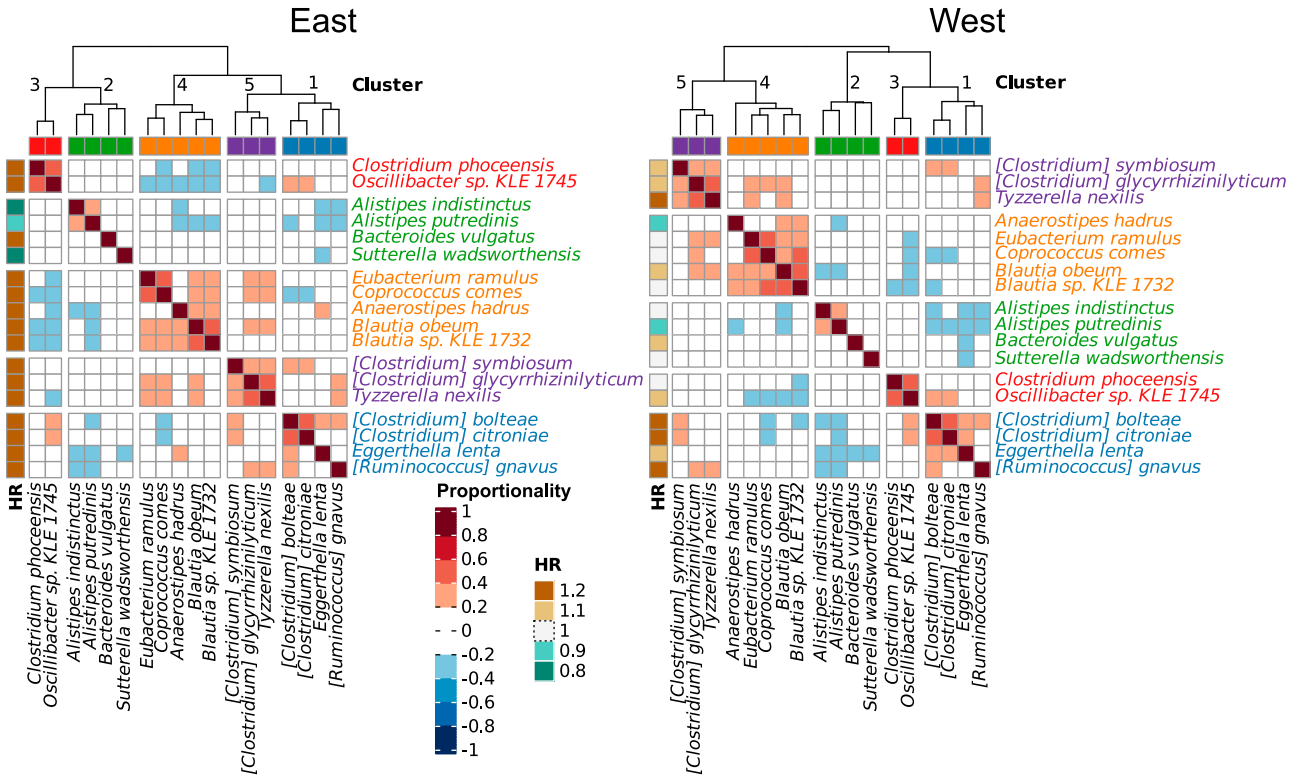

Figure 1—Proportionality between bacterial taxa significantly associated with incident type 2 diabetes in eastern Finland and western Finland. Annotated HRs and clustering of the taxa were calculated separately in both data groups. Because of identical cluster membership of the taxa, the cluster numbers and their annotations are harmonized.

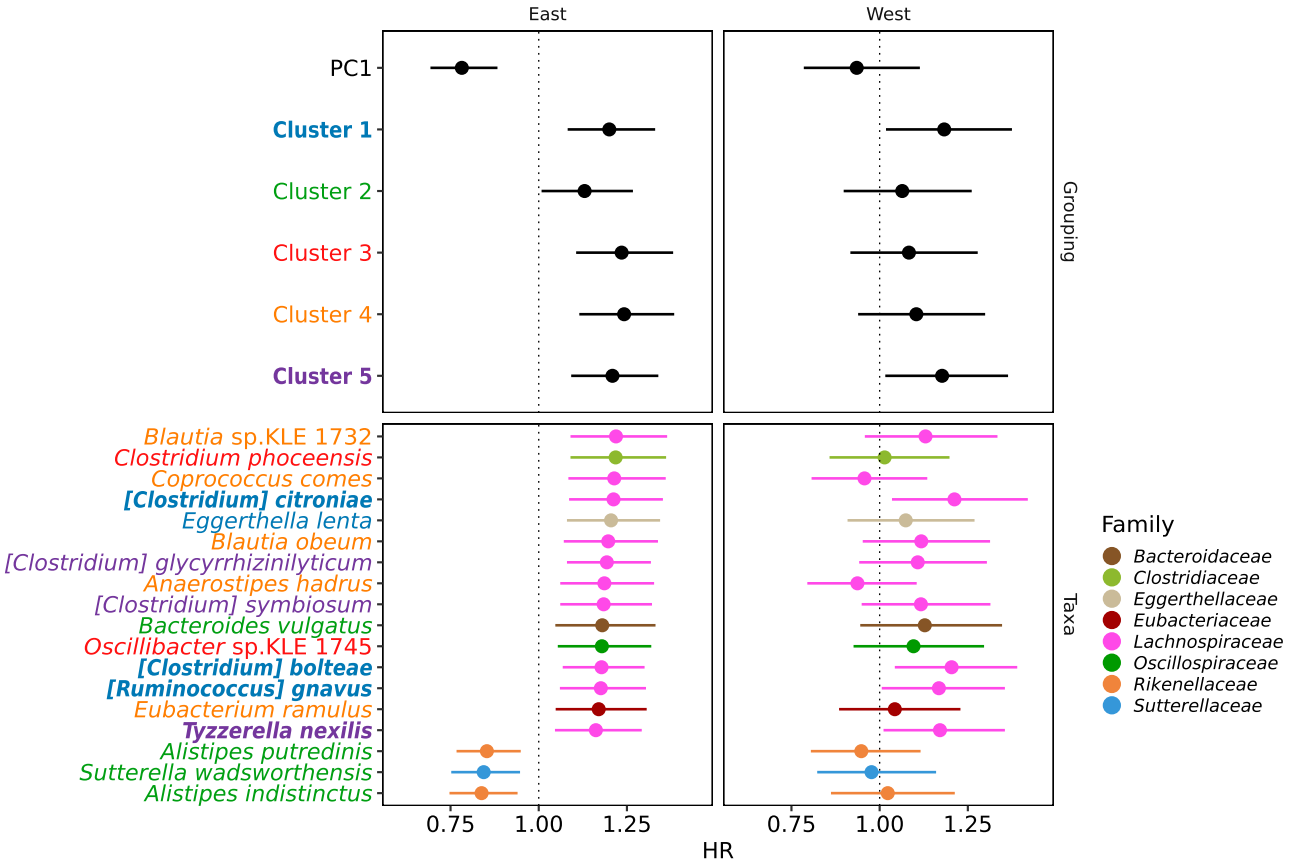

Figure 2—Comparison of HRs between models for the selected features in eastern and western Finland data. Features with significant associations in the validation (western Finland) data are indicated in bold, and the taxon colors show their membership in a cluster. The information in this figure can be found in numeric format in Supplementary Table 1.

groups, where taxon membership in each cluster was highly similar to that in the full data (Supplementary Fig. 1). Notably, cluster 4 in the full data was divided into clusters 4 and 5 in the subset data. The clustering pattern in the subset data also remained mostly robust between eastern and western Finland data. Briefly, two species changed cluster membership, and clusters 4 and 5 merged into a single cluster in the eastern Finland subset data, compared with the western Finland subset data (Supplementary Fig. 1). In the validation of the associations in the western Finland subset data, the same four species and clusters 1 and 6 were significantly associated with increased risk of diabetes as in the western Finland full data (unadjusted $P < 0.05$) (Supplementary Fig. 2 and Supplementary Table 2). Cluster 1 in the subset data included the same species as cluster 1 in the full data, together with *Dorea* sp. 5-2, which was associated with incident diabetes only in the eastern Finland subset data. Cluster 6 in the subset data had identical species membership to cluster 5 in the full data. Furthermore, the Kaplan-Meier survival curves of both the individual species and clusters in western Finland data showed similar trends to the corresponding features in the full data (Supplementary Fig. 3).

CONCLUSIONS

Previous studies have identified several biometric, genetic, and lifestyle risk factors for incident type 2 diabetes and established their role in its development (25). After adjusting for several known risk factors, we demonstrated that several common taxa in the gut microbiome among healthy Finnish adults were associated with incident type 2 diabetes over long-term follow-up. Specifically, we identified four species in the family *Lachnospiraceae* robustly associated with a higher type 2 diabetes risk in two geographically and genetically separate regions of Finland. Three of these taxa could be clustered together by proportional abundance in both geographic areas, and combined abundance of the four taxa was also predictive of incident type 2 diabetes.

Our findings are supported by several prior cross-sectional observations of microbiome composition related to type 2 diabetes and its risk factors. For example, *C. citroniae* has been positively associated with production of trimethylamine N-oxide (TMAO), which is a compound likely connected to intake of red meat (26). The direct association between red meat intake and type 2 diabetes risk has been known for >15 years (27). Furthermore, TMAO has been implicated in adipose tissue inflammation and impeded hepatic insulin signaling, which are connected to increased insulin resistance, high blood glucose levels, and type 2 diabetes (28). *C. bolteae* was reported to be enriched in patients with type 2 diabetes in a previous cross-sectional study along with other opportunistic pathogens (4). Interestingly, the abundance of *C. bolteae* was reduced in patients treated with acarbose, an α-glucosidase inhibitor used as an antidiabetic drug (29).

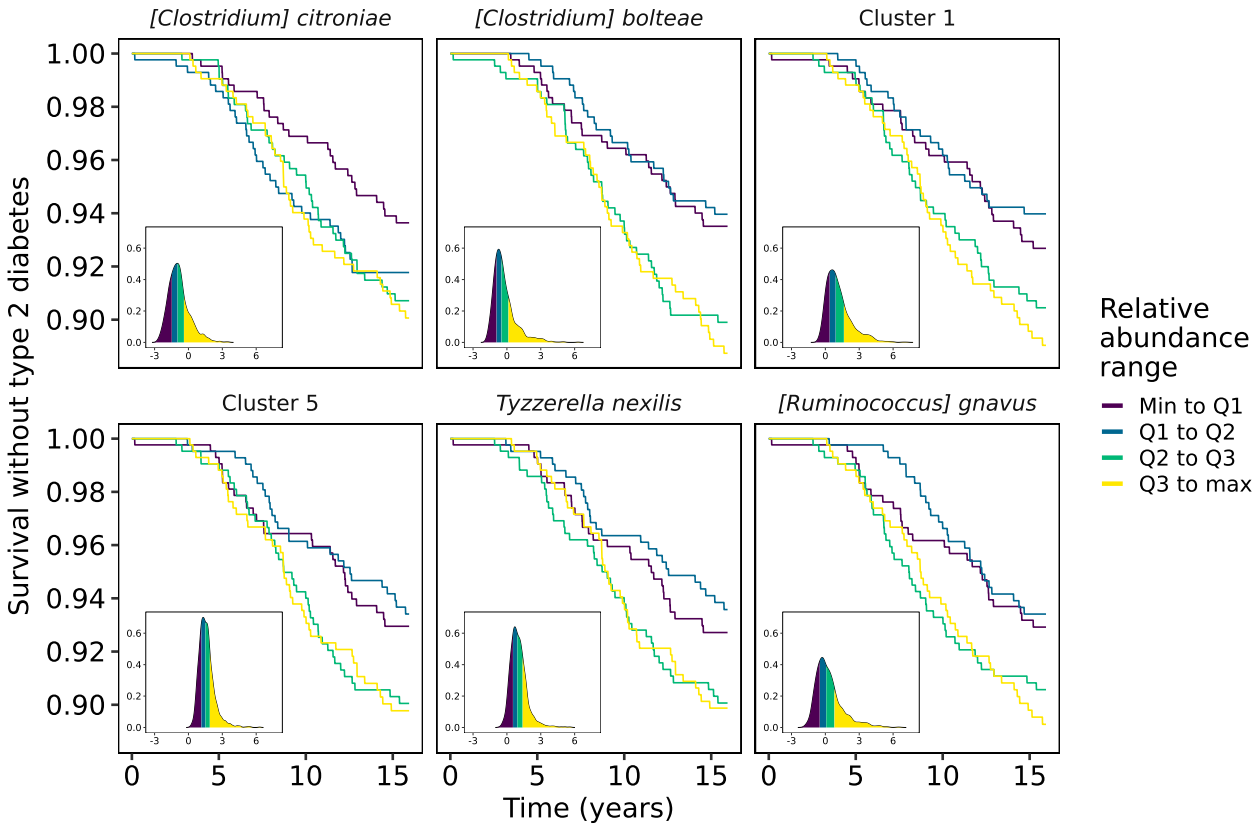

Figure 3—Kaplan-Meier curves for features with significant effect sizes in both data sets, displaying diabetes-free survival times of participants in western Finland. Curves are separated by ranges between quartiles of relative abundance of each feature. Distribution of the participants with the same relative abundance ranges is included as an inlay for each of the features.

Acarbose works by inhibiting the breakdown of complex polysaccharides in the small intestine, which makes these compounds available for microbes in the colon and helps to lower blood glucose levels through the slower uptake of simple sugars. Also, the abundance of *T. nexilis* has been observed to decrease drastically in response to intake of polydextrose, a soluble fiber (30). Polydextrose supplementation in connection with a high-fat diet has been reported to increase the concentration of postprandial plasma glucagon-like peptide-1, which is involved in regulation of blood glucose levels (31). The abundance of *C. bolteae* and *T. nexilis* appears to be related to intake and availability of different polysaccharides in the colon, which likely influences their ecologic niche. However, the mechanistic details of the link between these taxa and blood glucose levels remains to be clarified in detail. The abundance of *R. gnavus* is potentially related to glucose metabolism regulation and linked to increases in inflammatory cytokines, both of which are related to type 2 diabetes pathophysiology (5,32).

All four observed diabetes-associated taxa have been previously linked with other metabolic diseases and risk factors. For example, *R. gnavus* has been positively associated with obesity in animals (33,34) and humans (35). These taxa have also been associated with serum γ-glutamyl transferase levels, an important liver disease marker (36). Our previous cross-sectional study of fatty liver disease in FINRISK 2002 also features serum γ-glutamyl transferase level as a component of the modeled risk index and detected positive associations of all four taxa with higher disease risk (37). Thus, the results of the current study support several links between dietary habits, metabolic diseases, and type 2 diabetes, likely modulated by the gut microbiome.

While only some of the associations with individual gut microbiome taxa in eastern Finland were detected in western Finland, remarkably, the 18 taxa associated with type 2 diabetes in the East clustered identically in the West (Fig. 1). The association directions of the features with incident type 2 diabetes were mostly consistent between data from eastern and western Finland, as were features with statistically inconclusive results (Fig. 2). However, there were also several taxa with inconsistent association directions between the two data groups. The eastern and western Finland subpopulations had statistically significant differences in BMI, systolic blood pressure, non-HDL cholesterol, and blood glucose levels (all unadjusted $P < 0.05$) (Table 1). It is possible that these differences contributed to the inconsistencies in the microbial associations. However, the geographic distance between the regions and the differences in ethnic and lifestyle features of the subpopulations are more likely to have caused the partly inconsistent taxon associations (38).

All the robust positive associations of taxa and clusters were also robust after the exclusion of participants diagnosed

with type 2 diabetes within 2 years of follow-up. This result indicates that these microbial signals were likely associated with a long-term risk of developing type 2 diabetes and did not reflect changes in the gut microbiome caused by undiagnosed type 2 diabetes or its treatment (9).

The difference in type 2 diabetes incidence among the relative abundance quartiles of all robustly associated features emerged only after ~5 years of follow-up (Fig. 3). Therefore, it might have been challenging to detect taxon associations in previous studies with shorter follow-up times (7) or cross-sectional settings. Furthermore, the relative abundance distributions of all the features were slightly skewed, with long tails of higher values. For features other than *C. citroniae*, the long-term risk of incident type 2 diabetes, however, seemed to be increasing only after relative abundance values above the median. The relative abundance of *C. citroniae* was, however, quite low compared with the other taxa (or clusters), and only participants with relative abundances of this species below Q1 seemed to have a lower risk of developing incident type 2 diabetes. Thus, the metabolism of *C. citroniae*, including its potential for TMAO production (26,39), might be important for the pathogenesis of type 2 diabetes.

The two species historically classified in the genus *Clostridium* (*C. citroniae* and *C. bolteae*) have recently been reclassified into a new genus, *Enterocloster* (40). This close phylogenetic relatedness might further indicate sharing of metabolic traits between these taxa. Also, other members of this new genus, such as *C. clostridioforme*, have been associated with metabolic diseases, such as fatty liver disease (37), and with production of TMAO (26). Additionally, *C. clostridioforme* and *C. symbiosum* produce 3-methyl-4-(trimethylammonio)butanoate and 4-(trimethylammonio)pentanoate, which have been reported to be mechanistically linked to type 2 diabetes pathogenesis (41). These connections could warrant further study of the members of the new genus *Enterocloster* and their connections with chronic diseases. Furthermore, many of the taxon associations in our study have only been previously observed with shotgun metagenomics (4,26,29,30,36). For example, to our knowledge *C. bolteae* and *T. nexilis* have not been associated with type 2 diabetes in studies where 16S rRNA amplicon sequencing has been used. Studies reporting associations between type 2 diabetes and gut microbiome composition should thus preferably use, for example, full-length 16S rRNA gene sequencing or shotgun metagenomics instead of 16S rRNA metabarcoding. The construction of microbiome risk scores for type 2 diabetes is a promising approach to aid in its diagnosis and prevention (7). However, this method would also benefit from higher taxonomic resolution enabled by shotgun metagenomics or full-length 16S rRNA gene sequencing instead of amplicon sequencing.

The strengths of the current study include the high taxonomic coverage and resolution of shotgun sequencing, long follow-up time, and a large unselected study sample. Our results were also not confounded by antidiabetic drugs, including metformin. Therefore, the microbial signals we detected are more likely associated with type 2 diabetes progression or onset than related to the effects of dysglycemia (9). The associations were not affected by the exclusion of individuals with possibly undiagnosed type 2 diabetes (i.e., individuals who developed diabetes over the first years of follow-up). In addition, the detected microbial signals support several previous cross-sectional observations on connections between the gut microbiome and type 2 diabetes detected in different populations. The signals were also robust in the geographically and genetically distinct regions in Finland. The prevalence of type 2 diabetes in Finland is slightly higher than in other European countries on average (42). However, the demography and burden of risk factors in Finland are similar to those in other Nordic countries, which rank globally highly on a range of sociodemographic and health-relevant measures (43). Nevertheless, we acknowledge that these factors might affect the generalization of our results to other countries. Furthermore, the use of shallow shotgun metagenomics enables only a description of associations between taxa and incident disease, because the depth of the sequencing prevents genome assembly. Also, our incident type 2 diabetes definition combined both in- and outpatient disease diagnoses, drug prescriptions, and drug reimbursement data. Although the completeness and accuracy of these register data can be considered excellent (44), it is possible that some cases were not diagnosed during the follow-up period, especially at an early stage of disease progression.

We are not aware of previous long-term prospective studies of the associations between type 2 diabetes and the gut microbiome, similar to the current study. Therefore, our results should be further validated with studies in suitable cohorts to address their generalizability. Similar prospective studies with long follow-up times of >5 years can be a powerful tool to detect early signals of diseases with known connections to gut microbiome composition. Finally, we note that additional experiments in humans and animal models could likely establish the required mechanistic and causal evidence to link specific microbial species and strains conclusively to type 2 diabetes pathogenesis. The current study thus serves as a stepping stone toward the goal of improved prediction and the development of effective treatments for type 2 diabetes through modification of the gut microbiome.

Acknowledgments. The authors thank all participants in the FINRISK 2002 study and Tara Schwartz for assistance with laboratory work.

Funding. This research was supported in part by grants from the Finnish Cultural Foundation, the Finnish Foundation for Cardiovascular Research, the Emil Aaltonen Foundation, the Finnish Medical Foundation, the Sigrid Juselius Foundation, and the Academy of Finland (338818 [M.O.R.], 321356 [A.S.H.], 295741 and 307127 [L.L.], and 321351 [T.J.N.]).

Duality of Interest. Additional support was provided by Illumina, Inc., and Janssen Pharmaceutica through their sponsorship of the Center for Microbiome Innovation at University of California San Diego. T.J.N. has received speaking honoraria from Servier. V.S. has consulted for and received an honorarium from Sanofi and has an ongoing research collaboration with Bayer AG, all unrelated to this study. J.T. owns stocks in Orion Pharma and has received an honorarium from Eli Lilly. No other potential conflicts of interest relevant to this article were reported.

Author Contributions. M.O.R. and P.P.E. wrote the manuscript in consultation with all authors. M.O.R., P.P.E., V.S., R.K., L.L., and T.J.N. designed the work. M.O.R., P.P.E., L.L., and T.J.N. analyzed the data. A.S.H., G.M., J.T., P.J., V.S. and R.K. acquired the data. M.I., P.J., V.S., R.K., L.L., and T.J.N. supervised the work. Y.L. and M.J. provided critical feedback and suggestions on the draft versions of the manuscript. All authors gave final approval of the version to be published. T.J.N. is the guarantor of this work and,

as such, had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Prior Presentation. A non–peer-reviewed version of this article was submitted to the medRxiv preprint server (https://doi.org/10.1101/2021.11.10.21266163) on 11 November 2021.

References

1. Scott RA, Scott LJ, Mägi R, et al.; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium. An expanded genome-wide association study of type 2 diabetes in Europeans. Diabetes 2017;66:2888–2902
2. Patel CJ, Bhattacharya J, Butte AJ. An environment-wide association study (EWAS) on type 2 diabetes mellitus. PLoS One 2010;5:e10746
3. Li Q, Chang Y, Zhang K, Chen H, Tao S, Zhang Z. Implication of the gut microbiome composition of type 2 diabetic patients from northern China. Sci Rep 2020;10:5450
4. Qin J, Li Y, Cai Z, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 2012;490:55–60
5. Gurung M, Li Z, You H, et al. Role of gut microbiota in type 2 diabetes pathophysiology. EBioMedicine 2020;51:102590
6. Aw W, Fukuda S. Understanding the role of the gut ecosystem in diabetes mellitus. J Diabetes Investig 2018;9:5–12
7. Gou W, Ling CW, He Y, et al. Interpretable machine learning framework reveals robust gut microbiome features associated with type 2 diabetes. Diabetes Care 2021;44:358–366
8. Reitmeier S, Kiessling S, Clavel T, et al. Arrhythmic gut microbiome signatures predict risk of type 2 diabetes. Cell Host Microbe 2020;28:258–272.e6
9. Forslund K, Hildebrand F, Nielsen T, et al.; MetaHIT consortium. Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. Nature 2015;528:262–266
10. Wu H, Esteve E, Tremaroli V, et al. Metformin alters the gut microbiome of individuals with treatment-naive type 2 diabetes, contributing to the therapeutic effects of the drug. Nat Med 2017;23:850–858
11. Borodulin K, Tolonen H, Jousilahti P, et al. Cohort profile: the national FINRISK study. Int J Epidemiol 2018;47:696–696i
12. Salosensaari A, Laitinen V, Havulinna AS, et al. Taxonomic signatures of cause-specific mortality risk in human gut microbiome. Nat Commun 2021;12:2671
13. Marotz L, Schwartz T, Thompson L, et al. Earth Microbiome Project (EMP) high throughput (HTP) DNA extraction protocol v1 (protocols.io.pdmdi46). Accessed 10 November 2020. Available from https://www.protocols.io/view/earth-microbiome-project-emp-high-throughput-htp-d-pdmdi46
14. Sanders JG, Nurk S, Salido RA, et al. Optimizing sequencing protocols for leaderboard metagenomics by combining long and short reads. Genome Biol 2019;20:226
15. Glenn TC, Nilsen RA, Kieran TJ, et al. Adapterama I: universal stubs and primers for 384 unique dual-indexed or 147,456 combinatorially-indexed Illumina libraries (iTru & iNext). PeerJ 2019;7:e7755
16. Didion JP, Martin M, Collins FS. Atropos: specific, sensitive, and speedy trimming of sequencing reads. PeerJ 2017;5:e3720
17. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;9:357–359
18. Hillmann B, Al-Ghalith GA, Shields-Cutler RR, Zhu Q, Knight R, Knights D. SHOGUN: a modular, accurate, and scalable framework for microbiome quantification. Bioinformatics 2020;36:4088–4090
19. R Core Team. R: a language and environment for statistical computing. Vienna, Austria, R Foundation for Statistical Computing, 2019. Accessed 18 February 2020. Available from https://www.R-project.org/
20. Therneau T. A package for survival analysis in R. Accessed 4 August 2021. Available from https://CRAN.R-project.org/package=survival
21. Quinn TP, Richardson MF, Lovell D, Crowley TM. propr: an R-package for identifying proportionally abundant features using compositional data analysis. Sci Rep 2017;7:16252
22. White D, Gramacy R. maptree: mapping, pruning, and graphing tree models. 2012. Accessed 4 August 2021. Available from https://CRAN.R-project.org/package=maptree
23. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics 2016;32:2847–2849
24. Harrell FE Jr. rms: regression modeling strategies. 2021. Accessed 3 September 2021. Available from https://CRAN.R-project.org/package=rms
25. Zhang Y, Pan X-F, Chen J, et al. Combined lifestyle factors and risk of incident type 2 diabetes and prognosis among individuals with type 2 diabetes: a systematic review and meta-analysis of prospective cohort studies. Diabetologia 2020;63:21–33
26. Li J, Li Y, Ivey KL, et al. Interplay between diet and gut microbiome, and circulating concentrations of trimethylamine N-oxide: findings from a longitudinal cohort of US men. Gut. 29 April 2021. [Epub ahead of print]. DOI: https://doi.org/10.1136/gutjnl-2020-322473
27. Song Y, Manson JE, Buring JE, Liu S. A prospective study of red meat consumption and type 2 diabetes in middle-aged and elderly women: the Women's Health Study. Diabetes Care 2004;27:2108–2115
28. Shan Z, Sun T, Huang H, et al. Association between microbiota-dependent metabolite trimethylamine-*N*-oxide and type 2 diabetes. Am J Clin Nutr 2017;106:888–894
29. Gu Y, Wang X, Li J, et al. Analyses of gut microbiota and plasma bile acids enable stratification of patients for antidiabetic treatment. Nat Commun 2017;8:1785
30. Creswell R, Tan J, Leff JW, et al. High-resolution temporal profiling of the human gut microbiome reveals consistent and cascading alterations in response to dietary glycans. Genome Med 2020;12:59
31. Olli K, Salli K, Alhoniemi E, et al. Postprandial effects of polydextrose on satiety hormone responses and subjective feelings of appetite in obese participants. Nutr J 2015;14:2
32. Rodrigues RR, Gurung M, Li Z, et al. Transkingdom interactions between Lactobacilli and hepatic mitochondria attenuate western diet-induced diabetes. Nat Commun 2021;12:101
33. Petriz BA, Castro AP, Almeida JA, et al. Exercise induction of gut microbiota modifications in obese, non-obese and hypertensive rats. BMC Genomics 2014;15:511
34. Zheng X, Huang F, Zhao A, et al. Bile acid is a significant host factor shaping the gut microbiome of diet-induced obese mice. BMC Biol 2017;15:120
35. Verdam FJ, Fuentes S, de Jonge C, et al. Human intestinal microbiota composition is associated with local and systemic inflammation in obesity. Obesity (Silver Spring) 2013;21:E607–E615
36. Li R-J, Jie Z-Y, Feng Q, et al. Network of interactions between gut microbiome, host biomarkers, and urine metabolome in carotid atherosclerosis. Front Cell Infect Microbiol 2021;11:708088
37. Ruuskanen MO, Åberg F, Männistö V, et al. Links between gut microbiome composition and fatty liver disease in a large population sample. Gut Microbes 2021;13:1–22
38. Ruuskanen MO, Sommeria-Klein G, Havulinna AS, Niiranen TJ, Lahti L. Modelling spatial patterns in host-associated microbial communities. Environ Microbiol 2021;23:2374–2388
39. Martínez-del Campo A, Bodea S, Hamer HA, et al. Characterization and detection of a widely distributed gene cluster that predicts anaerobic choline utilization by human gut bacteria. MBio 2015;6:e00042-15
40. Haas KN, Blanchard JLY. Reclassification of the *Clostridium clostridioforme* and *Clostridium sphenoides* clades as *Enterocloster* gen. nov. and *Lacrimispora* gen. nov., including reclassification of 15 taxa. Int J Syst Evol Microbiol 2020;70:23–34
41. Ormsby MJ, Hulme H, Villar VH, et al. Microbiome-derived metabolites reproduce the mitochondrial dysfunction and decreased insulin sensitivity observed in type 2 diabetes. Accessed 21 October 2021. Available from https://www.biorxiv.org/content/10.1101/2020.08.02.232447v1
42. Bommer C, Sagalova V, Heesemann E, et al. Global economic burden of diabetes in adults: projections from 2015 to 2030. Diabetes Care 2018;41:963–970
43. Nordic Burden of Disease Collaborators. Life expectancy and disease burden in the Nordic countries: results from the Global Burden of Diseases, Injuries, and Risk Factors Study 2017. Lancet Public Health 2019;4:e658–e669
44. Sund R. Quality of the Finnish Hospital Discharge Register: a systematic review. Scand J Public Health 2012;40:505–515