Taylor & Francis
Taylor & Francis Group

RESEARCH PAPER

Check for updates

# FUT2 genotype and secretory status are not associated with fecal microbial composition and inferred function in healthy subjects

Williams Turpin [a,b], Larbi Bedrani[a,b], Osvaldo Espin-Garcia [c,d], Wei Xu[d], Mark S. Silverberg[a], Michelle I. Smith[a,b], David S. Guttman[e,f], Anne Griffiths[g], Paul Moayyedi[h], Remo Panaccione[i], Hien Huynh[j], Hillary Steinhart[a,b], Guy Aumais[k], Konstantin Shestopaloff [d], Levinus A. Dieleman[l], Dan Turner, CCC IBD GEM Project research team[m], Andrew D. Paterson[d,n,o], and Kenneth Croitoru [a,b]

[a]Zane Cohen Centre for Digestive Diseases, Mount Sinai Hospital, Toronto, ON, Canada; [b]Division of Gastroenterology, Department of Medicine, University of Toronto, Toronto, ON, Canada; [c]Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Ontario, Canada; [d]Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada; [e]Department of Cell & Systems Biology, University of Toronto, Toronto, Ontario, Canada; [f]Centre for the Analysis of Genome Evolution & Function, University of Toronto, Toronto, Ontario, Canada; [g]Division of Gastroenterology, Hepatology and Nutrition, Department of Paediatrics, The Hospital for Sick Children, Toronto, Ontario, Canada; [h]Department of Medicine, McMaster University, Hamilton, Ontario, Canada; [i]Inflammatory Bowel Disease Clinic, Division of Gastroenterology and Hepatology of Gastroenterology, University of Calgary, Calgary, Alberta, Canada; [j]Department of Pediatrics, University of Alberta, Edmonton, Alberta, Canada; [k]Montreal University, Hôpital Maisonneuve-Rosemont, Department of Medicine, Montreal, Quebec, Canada; [l]Division of Gastroenterology and CEGIIR, Department of Medicine, University of Alberta, Edmonton, Alberta, Canada; [m]Shaare Zedek Medical Center, Department of pediatric GI, Jerusalem, Israel; [n]Division of Epidemiology, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada; [o]Genetics and Genome Biology, The Hospital for Sick Children Research Institute, The Hospital for Sick Children, Toronto, Ontario, Canada

**ABSTRACT**

Heritability analysis of the microbiota has demonstrated the importance of host genotype in defining the human microbiota. The alpha (1,2)-fucosyltransferase 2 encoded by FUT2 is involved in the formation of the H antigen and the SNP, rs601338 is associated with ABO histo-blood group antigen secretion in the intestinal mucosa. Previous studies have provided non replicated results for the association of this polymorphism with the composition and inferred function of intestinal microbiota. We aimed to assess this relationship in a large cohort of 1,190 healthy individuals. Genotyping was performed using the HumanCoreEXOME chip, microbial composition was addressed by 16S rRNA gene sequencing. Firmicutes, Bacteroidetes, and Actinobacteria were the dominant phyla in this cohort. Although we have sufficient power to detect significant associations of FUT2 genotype/ inferred phenotype with the microbiota, our data demonstrate that FUT2 genotype and secretor status is not associated with microbial alpha diversity, microbial composition or inferred microbial function after correction for multiple testing. Thus, *FUT2* genotype and inferred phenotype are not associated with human fecal microbial composition and imputed function.

## Introduction

The human gut microbiome is involved in a number of important metabolic functions.[1] The major microbial phyla comprising the human gut microbiome are Firmicutes, Bacteroidetes, Actinobacteria, and Proteobacteria[2] At the genus taxonomy level, the composition varies markedly in healthy subjects, with some subjects having > 80% of their microbiota composition represented by a single taxa such as *Bacteroides* or *Prevotella* genera.[3] Although, the function and composition of the human microbiome are subject to long-term temporal stability there is considerable short-term variability.[4] The reason for high inter-individual variability of the composition of the microbiome in healthy individual is unknown; however, diet, environmental factors and host genetics likely contribute to the variability of the composition of the human gut microbiome.[4-6]

Heritability analysis of the microbiota has demonstrated the importance of host genotype in

---

**CONTACT** Dr. Kenneth Croitoru ✉ Kcroitoru@mtsinai.on.ca 🏢 Zane Cohen Centre for Digestive Diseases, Division of Gastroenterology, Department of Medicine, University of Toronto, Mount Sinai Hospital, 600 University Avenue Room 437, Toronto, Ontario, M5G 1 × 5, Canada.

🅑 Supplemental data for this article can be accessed on the publisher's website.

defining the human microbiota.[6,7] Although our data in a large cohort failed to show any association with inflammatory bowel diseases (IBD) risk associated genetic polymorphisms.[7] other studies suggest that specific genetic variants in key genes involved in mucus composition such as *FUT2* are associated with microbial composition.[8-11] Therefore, the association of IBD risk associated genetic polymorphisms with the diversity of the human microbiota remains controversial.[12,13] The protein encoded by the alpha (1,2)-fucosyltransferase 2 gene (*FUT2*) is responsible for secretion of the ABO histo-blood group antigens in the mucosa. The *FUT2* associated SNP rs601338 is associated with the secretor/non secretor status. The minor allele (A) confers a non-secretor phenotype and is associated with Crohn's disease (CD) susceptibility with odds ratio of 1.11 (1.071–1.143 95% CI, $p < 10^{-15}$).[13-15] Host-microbe interactions can involve *FUT2* through modification of components of the mucus layer,[16-19] and thus *FUT2* is an ideal candidate to test for association with the gut microbiota variability. Indeed several studies have already investigated the role of *FUT2* in the composition of the microbiota but large biological (mucosal versus stool sample or bile duct) and technical differences (DNA extraction and profiling method of the microbiota) exist between studies resulting in potentially different results.[8,10,11,20] Indeed the *FUT2* expression in tissue is the highest in salivary gland, then esophagus-mucosa, then terminal ileum, and then transverse colon[21] and it was shown that the effects of the *FUT2* genotype/secretor status on bacterial phyla decrease in distal region of the gastrointestinal tract.[22] In addition detecting an association of *FUT2* with microbial composition might also depend on other factor such as the stress applied to the host and/or diet of the host.[19,23]

We examined *FUT2* associated SNPs with intestinal microbial composition and inferred function using a cohort of subjects collected as part of a prospective cohort study of healthy first degree relatives of CD subjects (The Genetic, Environmental Microbial (GEM) project). We analyzed the association of *FUT2* with intestinal microbiota in 1,190 healthy first degree relatives of CD patients. Our results show that *FUT2* genotype and secreting status are not significantly associated with fecal microbiota composition and inferred function.

## Results

### Demographics of the cohort

The cohort comprised a total of 1,190 healthy individuals of European descent. The majority of the subjects were recruited from several Canadian provinces with the highest number from Ontario, Alberta and Quebec (37.3%, 24.5% and 14.4%) and 2.0% where recruited from either Israel or USA (Supplementary Table 1). In the studied cohort, there were more females (55.5%) and the mean age was 19.8 ± 7.7 (mean ± SD).

### Description of the microbiota composition of the cohort

The three dominant bacterial phyla were Firmicutes (relative abundance of 64.2% ± 14.1), Bacteroidetes (26.9% ± 15.0), and Actinobacteria (5.0% ± 5.2) (Supplementary Figure 2). Of the 127 genera, we found that *Blautia, Coprococcus, Ruminococcus, Bacteroides, Dorea, Roseburia, Faecalibacterium, Streptococcus* and *Oscillospira* genera were found in all subjects. The remaining 118 genera were irregularly observed as present or not present across subjects. For example two individuals harbour 8,468 and 4,881 reads assigned to a single OTU (28.2% and 16.2% of their microbiota) assigned to *Succinivibrio* genus, an OTU usually found in less than 5% of the general population. This confirms that our cohort have the existence of a core microbiome[24] coexisting with highly variable bacterial taxa of the human gut microbiome.[4]

### Association of rs516246 with the gut microbiota

When examined against all OTUs, rs516246 genotype and the inferred *FUT2* phenotype was not associated with microbial alpha diversity (Supplementary Tables 3, and Supplementary Figure 3). Because Bifidobacteriales taxon was previously associated with alpha diversity index we subsampled our microbiome data to perform an analysis restricted to this taxon.[9-11] Our data showed that Bifidobacteriales alpha diversity estimated by Chao1, Simpson, Shannon and observed species indices were not associated with *FUT2* genotype or secretor/non-secretor status (Supplementary Tables 4).

PCoA analysis of beta diversity performed on all OTUs as measured by unweighted UniFrac distances revealed that there was no clustering by genotype

($R^2 = 0.002$, p-value $= 0.31$) or inferred phenotype ($R^2 = 0.001$, p-value $= 0.18$) (Figure 1). Similar results was observed using Bray-Curtis distances across genotype ($R^2 = 0.002$, p-value $= 0.28$) or inferred phenotype ($R^2 = 0.001$, p-value $= 0.13$) (Figure 1). The difference in Bacteroidetes relative abundance reported previously among non-secretors was not replicated in our cohort (Figure 2).

We then applied the method published previously to analyze the microbiota at lower taxonomic levels.[8] Briefly, OTUs were filtered based on an observed presence in at least 60% of the samples and with a minimum total count of 30 across samples. This strategy reduces the total number of OTUs from 12,863 down to 396, and consequently reduces the number of comparisons tested. Again, *FUT2* genotype or inferred phenotype was not significantly associated with any of these filtered OTUs p>1.3 $10^{-4}$ (Supplementary table 5).[8]

### *FUT2 genotype/ inferred phenotype are not associated with microbial composition even with less stringent filtering of OTU table*

Filtering OTUs based on presence in at least 60% of the samples has important consequences on microbial composition characteristics (Supplementary Figure 1). A rarefaction at 30,000 reads, followed by filtering OTUs based on a presence in at least 60% of the samples decreased the number of reads to a mean of 23,952. For some individuals such filtering leads to an inaccurate estimation of the microbial composition (Supplementary Figure 1). By way of example, for five individuals, a single OTU assigned to *Prevotella copri* represents up to 40% of their microbiota. Filtering applied such that OTU's present in less than 60% of samples are removed would result in a finding that this bacteria represents <1% of the individual's entire fecal microbiota. In order to determine whether sensitivity to this filtering would alter the association between *FUT2* and microbiota, we repeated the association analysis to include OTUs with an observed presence in at least 5% of the samples. This strategy reduces the total number of OTUs from 12,863 to 4,353 without altering the overall microbial composition (Procrustes of OTU table with no filtering against OTU table filtered based on 5% of OTUs prevalence, $M^2 = 0.0$, p-value $<10^{-6}$) and removed most singleton and/or spurious OTUs.[25]

Given this compositional analysis we then assessed the associations between bacteria at the level of Phylum down to individual OTUs. Using this filtering of OTUs we were able to detect several nominal associations but they did not survive correction for multiple testing (Supplementary table 6).

### *Paired test in family members with* **FUT2** *inferred phenotype discrepancy*

Finally, we applied a different strategy to assess the relationship between secretor/non-secretor and microbial composition. Within the original cohort we had identified 102 individuals from 46 families who did not share the same secretor status (Secretor and non-secretor individuals from the same family). A paired t-test for related subjects failed to identify significant associations with taxa after correction for multiple testing (with the lowest p-value being 0.001 for Tissierellaceae taxa) (Supplementary table 7). The significant difference in phylotype abundances reported previously, namely, the increase in Bacteroidetes among non-secretors[8,9] was not confirmed in this analysis (Supplementary Figure 4).

### *Imputed function failed to detect any association between imputed microbial function and* **FUT2** *genotype and secretor/non-secretor status*

Microbial functions were imputed and tested for association with the *FUT2* genotype/inferred phenotype (See Material and Methods, Supplementary table 8, Supplementary table 9). It was previously shown that microbiome from secretor individuals separate from non-secretor.[8] Our PCoA analysis of beta diversity performed on all OTUs as measured by Bray-Curtis distances revealed that there was no clustering by inferred phenotype (R2 = 0.0004, p-value = 0.77).We did not identify any association of *FUT2* genotype or inferred phenotype with predicted metagenome composition by KEGG after correction of the p value for multiple testing (Supplementary table 9). Previous associations[8] in amino acid biosynthesis, in carbohydrate and lipid metabolism, in cofactors and vitamins metabolism and in glycan biosynthesis and metabolism in non-secretor (AA) individuals as compared to secretor (GG) were not replicated in our dataset. However some subcategories have nominal association with *FUT2* genotype such as K05988 (p-value <0.004) and K01799 (p-value <0.03) involved in carbohydrate
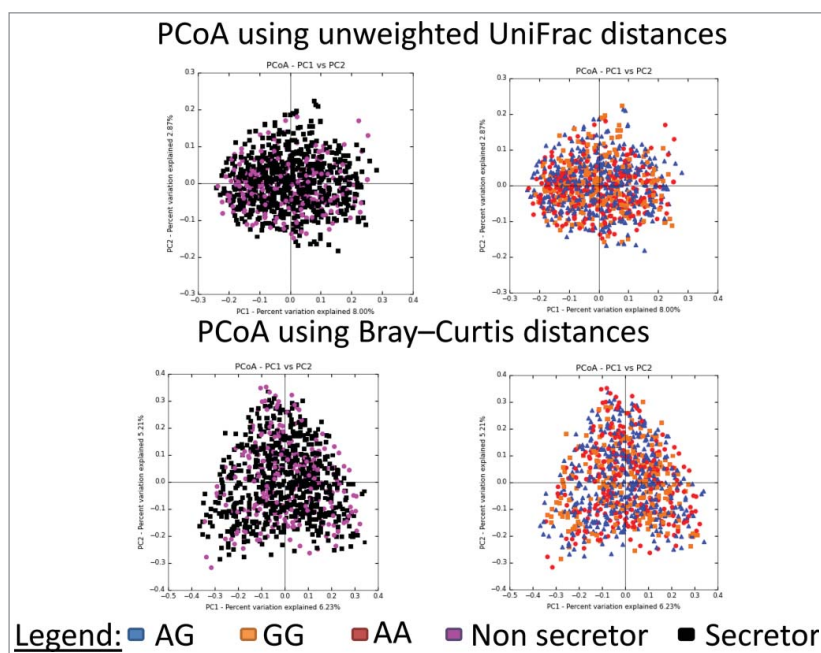
**Figure 1.** PCoA plot using Unweighted Unifrac and Bray-Curtis distances for beta diversity measure. Individuals who are either homozygous major (GG), heterozygous (GA) or homozygous minor (AA), at rs516246 are colored in orange, blue, and red respectively. Individuals with a secreting status are coloured in green and individuals non-secretor are coloured in purple.

metabolism, or K03918 (p-value <0.02) and K13831 (p-value <0.04) involved in amino acid metabolism (Supplementary table 9).

COG categories related to carbohydrate transport and metabolism (class [G]) were not significantly associated in pair comparison. Examining subcategories of COG specifically related to fucose metabolism, *i.e.* COG2407 (L-fucose isomerase and related proteins), COG3594 (Fucose 4-O-acetylase and related acetyltransferases), COG0738 (Fucose permease), and COG3669 (Alpha-L-fucosidase) failed to identify a significant difference across genotype groups. However subcategory COG4154 (Fucose dissimilation pathway protein FucU) had a nominal association with *FUT2* genotype (p-value <0.03) but this did not survive correction for multiple testing. Other COG function were nominally significant but were not closely link to fucose metabolism of the mucus (COG3465, an uncharacterized protein YwgA COG3676, a transposase and inactivated derivatives, COG3728, a phage terminase, and COG3774, a mannosyltransferase).

## Discussion

We examined a cohort of healthy first degree relatives of CD patients to assess if *FUT2*

polymorphism and inferred phenotype was associated with microbial composition and inferred function using 16S rRNA gene sequencing and PICRUSt software analysis to infer bacterial community function. The size of the cohort we examined allowed for the study to be sufficiently powered to detect the previously reported effect size for the association of *FUT2* with microbial composition.[8-11] Indeed, standard procedure for genetic-traits association was carefully applied that include, control of population stratification by restricting analysis to individuals of European descent, control of type I error by excluding related subjects (n = 272 individuals). With a total number of 918 unrelated individual, the power calculations in our study showed that we have >80% power to detect the association of *FUT2* rs516246 associations with 10 out of 13 bacterial phyla (Supplementary Tables 10) based on prior studies effect size.[8,9] Despite the fact that the size of our cohort is larger than that of previous studies,[8-11] our results indicated that *FUT2* is not associated with fecal microbial diversity, composition or inferred function. However, our results are in agreement with another recent study using a large cohort of 1,503 twins from the United Kingdom,[26] as well in 3 different microbiome GWAS
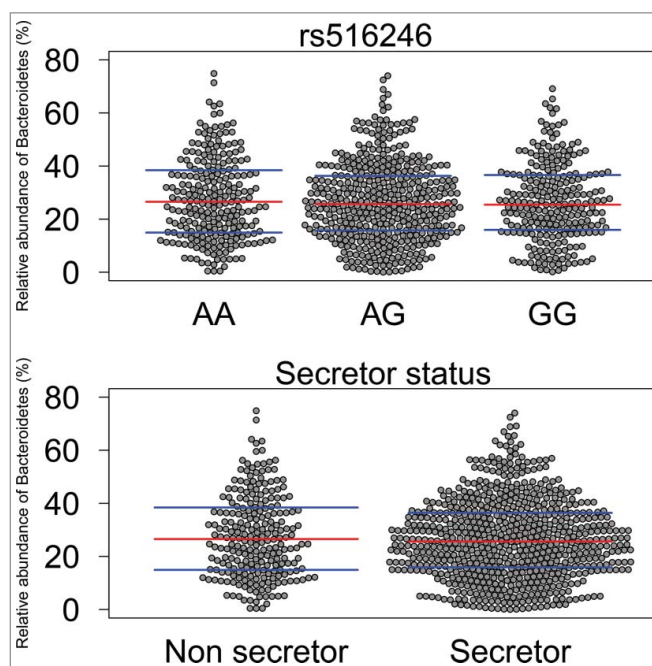
**Figure 2.** Beeswarm plot of the relative abundance of Bacteroidetes in individuals who are either homozygous major (GG), heterozygous (GA) or homozygous minor (AA), at rs516246 in 918 unrelated individuals of European descent. The lines represent the first (blue), second (bold red), and third quartiles (blue). Circles represent the relative abundance of the given family from an individual subject's sample.

studies that all failed to identify any association of FUT2 and fecal microbiome composition, even after adjusting for age, sex covariates and taking into account family structure.[7,27,28]

Alpha diversity was not significantly different across genotype using Chao1 and number of species index. These two indexes were applied and previously reported to be decreased in AA and GA as compared to GG, with a reported p values of 0.012 and 0.085 respectively.[8] However, another study using PCR Denaturing Gradient Gel Elecrophoresis (DGGE) and pyrosequencing to characterise microbial diversity[11] also failed to find any differences in alpha diversity with respect to FUT2 genotype. In addition, the same group found different results between DGGE, pyrosequencing and HITChip methods applied to the exact same cohort.[10,11] After subsampling our dataset to this taxon, we failed to replicate any association with Bifidobacteriales alpha diversity. Our sequencing protocol was directed at V4 region of the 16S which a region known to have adequate amplification and assignment of members of Bifidobacteriaceae family.[29-31] Thus, we believe that our results concerning Bifidobacteriaceaceare are robust.

We found that FUT2 genotype and inferred phenotype was not associated with microbial composition. This is different from previous studies which have generally been small in sample size.[8-11] Using PCR-DGGE fingerprinting technology to characterise the microbiome, it has been shown that non-secretors have lower relative abundance and diversity of the bifidobacterial taxa.[10,11] Wacklin et al. reported that this association was weak and pyrosequencing and HITChip characterization of the microbiota did not replicate all these findings in the same cohort.[10,11] Another study has shown that FUT2 secretor status was associated with gut microbiota compositio;n[9] however, only 47 individuals were studied including 29 with CD, which could indicate a confounding effect of disease phenotype rather than genotype.[9] The association of FUT2 with intestinal microbial the composition and function in the cecum and sigmoid colon from 39 healthy subjects has been reported. However these authors only included OTUs observed in at least 60% of the cohort reducing the number of OTUs from 4074 to 419. Here, a q-value threshold of 0.25 for a part of their analysis was used rather than the conventional p-value threshold of 0.05, thus increasing type I

error.[8] For comparison, we applied the same strategy, but we could still not replicate previous results.[8] Finally a small cohort of 35 individuals comprising 8 non-secretor individuals found no difference in alpha diversity and beta diversity but differences in Prevotellaceae, Paraprevotellaceae were observed.[32] Finally, a more recent large cohort study also found that secretor statuses are not associated with stool microbiota composition in 1,500 twins.[26] Our results are thus in accordance with this larger cohort study and suggest that studies with small sample size should be replicated in well powered studies.

FUT2 genotype/ inferred phenotype was not associated with imputed microbial function assessed using PICRUSt. A report from Tong et al. was the only previous study to evaluate the function of the human microbiota in the context of FUT2.[8] FUT2 was shown to be associated with several KEGG pathways including carbohydrate and lipid metabolism, cofactors and vitamins metabolism and glycan biosynthesis and metabolism. In our dataset none of these pathways associations were replicated. The discordance between our report and that of Tong et al could be due to numerous technical differences: DNA extraction (PowerSoil DNA Isolation Kit versus QIAamp DNA Stool Mini Kit), sequencing technology (HiSeq versus MiSeq). We used the same primers to sequence the 16S gene; however, HiSeq allows the generation of a higher numbers of reads per run compared to MiSeq.[30] Due to lower number of sequences generated by MiSeq we chose a rarefaction of 30,000 sequences per sample while Tong et al. used a rarefaction of 300,000 sequences per sample. This difference in rarefaction depth might affect the presence/absence of particular OTUs with more rare OTUs and related inferred functions detected. As PICRUSt software uses OTU composition to infer bacterial function, it might be sensitive to OTUs richness and prevalence resulting in different conclusion.[33] A report by Langille et al. has demonstrated that only 105 16S sequences are required to accurately impute bacterial function.[34] Thus we believe that our conclusions are robust even with a lower number of sequences per sample.

The major difference between our study and previous reports is the bio-specimen used to investigate intestinal microbial composition. Tong et al. used mucosal lavage collected from the cecum and sigmoid colon.[8] Rash et al. used colonic sigmoid biopsies,[9] while in our study faecal samples were used. The microbial composition is known to differ across luminal and fecal samples.[35] Also, there is a decreasing gradient of fucose and ABH blood group expression from ileum to rectum.[36,37] Thus the alpha (1,2) fucosylated components are likely decreased in stool as compared to sigmoid or cecal samples. In addition, age is an important factor that is associated with differences in microbial composition.[38] The mean age of the GEM cohort (19.8 ± 7.7) was lower than previous studies[8,10,11] and thus the microbiota composition and/or inferred function in our population might be different from these studies. Nevertheless, the overall microbial composition in our cohort was consistent with that described in microbiome studies[2,6] (Supplementary Figure 2). Finally technical differences might explain some discrepancy such as DNA extraction, microbial identification method, primers, and regions of 16Sr RNA.[39,40] Indeed, DNA extraction was different across all these study and include the FastDNAH SPIN Kit for Soil (MP Biomedicals),[10,11] prep DNA/RNA Mini Kit (Qiagen),[9] and PowerSoil DNA Isolation Kit (MO BIO Laboratories)[8] while the QIAamp DNA Stool Mini Kit (Qiagen) was used in this study. These DNA extractions are known to result in differences in microbiome profiling even if the same sample was used for each extraction.[39] In addition, Waclkin et al. use DGGE, pyrosequencing and HITChip methods, while Rausch et al. used pyrosequencing and Tong et al. used Illumina HiSeq 2000 while Miseq was used in the GEM project.[41] All these methods make it difficult to directly compare the biological output of each study. To summarise, stool sampling, DNA extraction and the age of the cohort are the three most important factors which differ between this and previous studies which could explain the absence of a FUT2 association with microbial composition and inferred function observed in this study.

With 1,190 individuals, this study represents one of the largest human cohorts to assess the association of FUT2 genotype and inferred phenotype with the human gut microbiota. We found that FUT2 was nominally associated with numerous inferred microbial functions and with microbial composition but none of these associations survive correction of the p-value for multiple testing. In addition no clustering based on FUT2 genotype could be observed and the alpha diversity was similar across FUT2 genotype and secreting status. However dietary habits, variation in stool sampling, and DNA extraction might explain

discrepancy with other studies. In conclusion we found that *FUT2* was not associated with human microbiota. Genetic factors that are associated with microbial composition and function remain to be defined and replicated.[7,27,28]

## Methods

### Patient recruitment

Subjects were recruited as part of a prospective cohort study of healthy first degree relatives of CD subjects between 6 – 35 years of age from 2008 to 2013 (The Genetic, Environmental Microbial (GEM) project) (http://www.gemproject.ca/, Table 1, Supplementary Table 1). After consent, each subject provided stool and blood samples and completed a standardized questionnaire to exclude any history or symptoms of IBD or gastro-intestinal disease as defined by the clinical sub-committee of the GEM Project (Supplementary Note 1). Demographic information and environmental risk data were recorded (Supplementary Note 1, Supplementary Note 2, Table 1, and Supplementary Table 1). All subjects and/or their guardians gave written informed consent to participate in the study. Subjects were excluded if they had received antibiotic treatment within three months prior to their recruitment into the study. The study cohort comprised 1,190 asymptomatic individuals of European descent. The study was approved by the Mount Sinai Hospital Research Ethics Board (Toronto – Managing center) and each participating recruitment center.

### Genotyping of healthy first degree relatives by HumanCoreEXOME chip

Blood was collected from subjects and genomic deoxyribonucleic acid (DNA) extracted using the Gentra Puregene Blood Kit (Quiagen, CA, USA, catalog #158389). DNA samples were quantified by Nanodrop dilutions. DNA was then prepared at a final concentrations of 20 ng/$\mu$l and aliquoted into 96-well reaction plates. SNP genotyping was performed using the HumanCoreExome-24 v1.0 chip (Illumina, Inc. San Diego, CA). We restricted the analysis to subjects with self-declared Caucasian ethnicity due to the fact that rs601338 (W143X, G428A) is the most common cause of secretor status in European ancestry, while other polymorphisms may be responsible in other ethnicity

as reviewed previously.[42] Since, rs601338 is not genotyped in the HumanCoreEXOME chip we used rs516246, which is in strong linkage disequilibrium with rs601338 to infer secretor phenotype. According to the International HapMap Project (HapMap3 release 2, Northern and Western Ancestry population) these two SNPs are in perfect linkage disequilibrium with one another ($r^2 = 1.0$).[43] The minor allele (A) frequency was 49.9% and the genotype distribution were in Hardy-Weinberg equilibrium (p-value = 0.27). Individuals with AA genotype were defined as non-secretor while AG and GG were defined as secretor. Genotyping data and corresponding stool identifiers is available in Supplementary Table 11.

### Taxonomic profiling of the gut microbiota

Stool sampling from 1,190 subjects (comprising 918 unrelated subjects) was performed as described previously.[41] Briefly, stool samples were collected in FB Commode Specimen Collector (Fisher Scientific, Waltham, MA, catalog #23-038032), put into Polypropylene vials (Starplex Scientific Inc, Etobicoke, ON, catalog #V302-F) and kept in freezers, then shipped from study sites and stored at -80°C. Fecal bacterial DNA was extracted using the QIAamp DNA Stool Mini Kit (Qiagen, Hilden, Germany, catalog #51504) with slight modifications which included physical disruption of the bacterial cell wall using ceramic beads. The V4 hypervariable region of bacterial 16S ribosomal RNA (16S) was sequenced in paired-end mode (2 × 150 base pair) on MiSeq platform (Illumina Inc., San Diego, CA, USA).[30] The resulting paired reads were assembled using PANDAseq v 2.7 to generate an amplicon size of 250 base pairs.[44] To test for inter and intra-sequencing run variation, a subset of 45 samples belonging to 12 different subjects were replicated (2–8 replicates depending on subjects) for quality control (Supplementary Table 2). Sequencing across multiple MiSeq runs was consistent (Supplementary Table 2). Assembled reads were demultiplexed and processed by the quantitative insights into microbial ecology (QIIME v1.8.0) pipeline using the default parameters.[45] Chimeric sequences were identified *de novo* and reference based and then removed using usearch61.[46] The non-chimeric sequences were then clustered into operational taxonomic units (OTUs) at 97.0% sequence similarity using a closed reference-based picking approach with UCLUST software

against Greengenes database 13_8 of bacterial 16S sequences.[47] All samples had a minimum of 30,000 reads after quality filtering. Alpha diversity was determined after rarefaction at a depth of 30,000 reads per sample using Chao1, Simpson, Shannon and observed species indices.[48]

## Power calculation

Power calculation was performed using QUANTO v1.2.4.[49] Power calculation was defined as the power at two-sided alpha of 5% to detect difference between genotypes using a recessive model, the observed coefficients of determination ($R^2$), and observed minor allele frequency in 918 unrelated individuals.

## Association of gut microbiota with FUT2 *genotype and inferred phenotype*

Since including related subjects in genetic association analysis inflates type 1 error,[50] we choose to restrict the first part of this study to 918 unrelated individuals of European descent. To assess if *FUT2* genotype/ inferred phenotype is associated with the microbial composition we applied the same methods as Tong et al.[8] which showed that rs516246 genotype (AG and GG vs AA) is associated with alpha diversity measured by Chao1 index and beta diversity estimated by using Bray-Curtis and unweighted UniFrac distances between samples.[8,51] In our cohort, we found that rs516246 genotype (AG and GG vs AA) have a similar alpha diversity. In addition, because Bifidobacteriales taxon was also previously associated with alpha diversity as determined by DGGE,[10] we restricted our microbiome dataset and investigated this taxon specifically. Bacterial community grouping was assessed using the Adonis test for 10,000 permutations on the first five dimensions of the principal coordinates analysis (PCoA).[52] Pair-wise comparisons between *FUT2* secretor versus non secretor was assessed using the non-parametric Kruskal-Wallis test to identify

differentially abundant phylum taxa and OTU levels of taxonomy.

Analysis of bacterial taxa and OTUs were performed using two different filtering. The first analysis applied the filtering used in Tong et al.[8] Briefly, to analyze at lower taxonomic levels, we filtered out low-abundant OTUs based on the criteria of minimum total observation count of 30 across all samples and being observed in at least 60% of the samples, reducing the total number of OTUs from 12,863 to 396.[8] A second approach was then applied because filtering OTU that are observed in at least 60% of the samples had consequences on characterization of each individual's microbiota (Procrustes of OTU table with no filtering against OTU table filtered based on 60% of OTUs prevalence, $M^2 = 0.01$, p-value $<10^{-6}$) with some individuals having a biased microbiota profiles after applying this filtering (Supplementary Figure 1– 2). Such approaches also decrease the number of comparison made. For such reason, we decided to apply less stringent criteria to filter OTUs that are observed in at least 5% of the samples and a minimum read count of 30 across all samples, leaving 4,353 OTUs (R software v2.14.1; http://CRAN.R-project.org). We then analysed bacterial composition based on OTUs grouped within the same taxonomic assignment at each level of bacterial taxonomy. A Bonferroni corrected p-value threshold of 0.05 was considered as significant (while 396 comparisons were performed at OTU level of bacterial taxonomy and 166 bacterial taxa). Raw p-values are reported. All genotype classes were compared in addition to secretor status and the p-value were corrected for intra-group compared.

## Imputed function of the gut microbiota

The function of the fecal microbial communities was imputed using PICRUSt V0.1.[34] Briefly, The OTU table was used as the input file for metagenome imputation after a rarefaction step to 30,000 sequences per sample. The pre-calculated table of gene counts was used to identify the gene counts in the organisms present in the microbiome.[53] The Kyoto Encyclopedia of Genes and Genomes (KEGG)[54] and clusters of orthologous groups (COG)[55] databases were used to identify the gene families. The metagenomic prediction had a relatively low weighted nearest sequenced taxon index (0.07 $\pm$. 0.017). Pair-wise comparisons between secretor/non-secretor were assessed using the non-

**Table 1.** Demographic and genetic data in unrelated individuals of European descent.

| | Total number of unrelated Subject | FUT2 genotype rs516246 | | |
|---|---|---|---|---|
| | (918) | GG (244) | AG (436) | AA (238) |
| *Gender* | F (501) | 119 | 246 | 136 |
| | M (417) | 125 | 190 | 102 |
| *Age* | Mean | 19.4 | 19.7 | 20.8 |
| | St Dev | 8.0 | 7.4 | 7.8 |

parametric Kruskal-Wallis test to identify differentially abundant inferred functions from KEGG and COG pathways. A Bonferroni correction for multiple testing was applied and significant association was considered below a corrected p-value threshold of 5%. All Procrust analyses were based on Bray-Curtis distance.[56] Monte Carlo simulation was performed using 10,000 permutations on the first five dimensions of the PCoA.

## Abbreviations

| | |
|---|---|
| CD | Crohn's disease |
| COG | clusters of orthologous groups |
| DGGE | Denaturing Gradient Gel Elecrophoresis |
| IBD | inflammatory bowel disease |
| FUT2 | alpha (1,2)-fucosyltransferase 2 |
| GEM | The Genetic, Environmental Microbial project |
| KEGG | The Kyoto Encyclopedia of Genes and Genomes |
| OTU | operational taxonomic units (OTUs) |
| PANDAseq | PAired-eND Assembler for DNA sequences |
| PICRUSt | Phylogenetic Investigation of Communities by Reconstruction of Unobserved States |
| PCoA | Principal Coordinates Analysis |
| QIIME | Quantitative Insights Into Microbial Ecology |
| SNP | single nucleotide polymorphism |

## Declarations

### Ethics approval and consent to participate
All subjects and/or their guardians gave written informed consent to participate in the study. The study was approved by the Mount Sinai Hospital Research Ethics Board (Toronto – Managing center) and local centers (07-0322-E).

### Consent for publication
Not applicable.

### Competing interests
All authors disclose no potential conflicts (financial, professional, or personal) that are relevant to the manuscript.

### Authors in the CCC IBD GEM project research team
The GEM Project Research Consortium is composed of: Maria Abreu, Paul Beck, Charles Bernstein, Kenneth Croitoru, Leo Dieleman, Brian Feagan, Anne Griffiths, David Guttman, Kevan Jacobson, Gilaad Kaplan, Denis O. Krause[*], Karen Madsen, John Marshall, Paul Moayyedi, Mark Ropeleski, Ernest Seidman, Mark Silverberg, Scott Snapper, Andy Stadnyk, Hilary Steinhart, Michael Surette, Dan Turner, Tom Walters, Bruce Vallance, Guy Aumais, Alain Bitton, Maria Cino, Jeff Critch, Lee Denson, Colette Deslandres, Wael El-Matary, Hans Herfarth, Peter Higgins, Hien Huynh, Jeff Hyams, David Mack, Jerry McGrath, Anthony Otley, and Remo Panancionne. ([*] deceased)

### Availability of data and materials
16S sequences are available the study accession number PRJEB14839 (NCBI BioProject database).

## Author contributions

Contributed equally – ADP and KC; Jointly supervised research- ADP and KC; Conceived and designed the experiments- WT, MSS, ADP, KC and The GEM Project Steering Committee; Performed the experiments- WT, WX, LX, KC, Performed statistical analysis; – WT, OEG, WX, LX, ADP; Analyzed the data – WT, LB, MSS, MIS, WX, GMH, DK, KS, OEG, DSG, LX, ADP, KC; Contributed reagents/materials/analysis tools and significant subject recruitment – WX, LX, ADP, AG, RP, AO, and the GEM Project Consortium; Wrote the paper- WT, ADP, KC.

## ORCID

Williams Turpin http://orcid.org/0000-0001-9364-5868

Osvaldo Espin-Garcia   ID   http://orcid.org/0000-0003-2052-2626

Konstantin Shestopaloff   ID   http://orcid.org/0000-0003-3933-4967

Kenneth Croitoru   ID   http://orcid.org/0000-0003-1231-0180

## References

1. Sanz Y, Olivares M, Moya-Perez A, Agostoni C. Understanding the role of gut microbiome in metabolic disease risk. Pediatric research. 2015;77:236–44. doi:10.1038/pr.2014.170. PMID:25314581.

2. Consortium THMP. Structure, Function and Diversity af the Healthy Human Microbiome. Nature. 2012;486:207–14. doi:10.1038/nature11234. PMID:22699609.

3. Gorvitovskaia A, Holmes SP, Huse SM. Interpreting Prevotella and Bacteroides as biomarkers of diet and lifestyle. Microbiome. 2016;4:15. doi:10.1186/s40168-016-0160-7. PMID:27068581.

4. Schloissnig S, Arumugam M, Sunagawa S, Mitreva M, Tap J, Zhu A, Waller A, Mende DR, Kultima JR, Martin J. Genomic variation landscape of the human gut microbiome. Nature. 2013;493:45–50. doi:10.1038/nature11711. PMID:23222524.

5. De Filippo C, Cavalieri D, Di Paola M, Ramazzotti M, Poullet JB, Massart S, Collini S, Pieraccini G, Lionetti P. Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. Proceedings of the National Academy of Sciences of the United States of America. 2010;107:14691–6. doi:10.1073/pnas.1005963107. PMID:20679230.

6. Goodrich JK, Waters JL, Poole AC, Sutter JL, Koren O, Blekhman R, et al. Human genetics shape the gut microbiome. Cell. 2014;159:789–99. doi:10.1016/j.cell.2014.09.053. PMID:25417156.

7. Turpin W, Espin-Garcia O, Xu W, Silverberg MS, Kevans D, Smith MI, Guttman DS, Griffiths A, Panaccione R, Otley A. Association of host genome with intestinal microbial composition in a large healthy cohort. Nature genetics. 2016;48:1413–7. doi:10.1038/ng.3693. PMID:27694960.

8. Tong M, McHardy I, Ruegger P, Goudarzi M, Kashyap PC, Haritunians T, Li X, Graeber TG, Schwager E, Huttenhower C. Reprograming of Gut Microbiome Energy Metabolism by the Fut2 Crohn's Disease Risk Polymorphism. The ISME journal. 2014;8:2193–206. doi:10.1038/ismej.2014.64. PMID:24781901.

9. Rausch P, Rehman A, Kunzel S, Hasler R, Ott SJ, Schreiber S, Rosenstiel P, Franke A, Baines JF. Colonic mucosa-associated microbiota is influenced by an interaction of Crohn disease and FUT2 (Secretor) genotype. Proceedings of the National Academy of Sciences of the United States of America. 2011;108:19030–5. doi:10.1073/pnas.1106408108. PMID:22068912.

10. Wacklin P, Makivuokko H, Alakulppi N, Nikkila J, Tenkanen H, Rabina J, Partanen J, Aranko K, Mättö J. Secretor genotype (FUT2 gene) is strongly associated with the composition of Bifidobacteria in the human intestine. PloS one. 2011;6:e20113. doi:10.1371/journal.pone.0020113. PMID:21625510.

11. Wacklin P, Tuimala J, Nikkila J, Sebastian T, Makivuokko H, Alakulppi N, Laine P, Rajilic-Stojanovic M, Paulin L, de Vos WM. Faecal microbiota composition in adults is associated with the FUT2 gene determining the secretor status. PloS one. 2014;9:e94863. doi:10.1371/journal.pone.0094863. PMID:24733310.

12. Cleynen I, Boucher G, Jostins L, Schumm LP, Zeissig S, Ahmad T, Andersen V, Andrews JM, Annese V, Brand S. Inherited determinants of Crohn's disease and ulcerative colitis phenotypes: a genetic association study. Lancet (London, England). 2016;387:156–67. doi:10.1016/S0140-6736(15)00465-1. PMID:26490195.

13. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, Lee JC, Schumm LP, Sharma Y, Anderson CA. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012;491:119–24. doi:10.1038/nature11582. PMID:23128233.

14. McGovern DP, Jones MR, Taylor KD, Marciante K, Yan X, Dubinsky M, Ippoliti A, Vasiliauskas E, Berel D, Derkowski C. Fucosyltransferase 2 (FUT2) non-secretor status is associated with Crohn's disease. Human molecular genetics. 2010;19:3468–76. doi:10.1093/hmg/ddq248. PMID:20570966.

15. Wu H, Sun L, Lin DP, Shao XX, Xia SL, Lv M. Association of Fucosyltransferase 2 Gene Polymorphisms with Inflammatory Bowel Disease in Patients from Southeast China. Gastroenterology research and practice. 2017;2017:4148651. doi:10.1155/2017/4148651. PMID:28167958.

16. Bry L, Falk PG, Midtvedt T, Gordon JI. A model of host-microbial interactions in an open mammalian ecosystem. Science (New York, NY). 1996;273:1380–3. doi:10.1126/science.273.5280.1380.

17. Meng D, Newburg DS, Young C, Baker A, Tonkonogy SL, Sartor RB, Walker WA, Nanthakumar NN. Bacterial symbionts induce a FUT2-dependent fucosylated niche on colonic epithelium via ERK and JNK signaling. American journal of physiology Gastrointestinal and liver physiology. 2007;293:G780–7. doi:10.1152/ajpgi.00010.2007. PMID:17673542.

18. Pham Tu, Anh N, Clare S, Goulding D, Arasteh Julia M, Stares Mark D, Browne Hilary P, Page AJ, Kumasaka N, Kane L. Epithelial IL-22RA1-Mediated Fucosylation Promotes Intestinal Colonization Resistance to an Opportunistic Pathogen. Cell Host & Microbe. 2014;16:504–16. doi:10.1016/j.chom.2014.08.017.

19. Pickard JM, Maurice CF, Kinnebrew MA, Abt MC, Schenten D, Golovkina TV, Bogatyrev SR, Ismagilov RF, Pamer EG, Turnbaugh PJ. Rapid fucosylation of intestinal epithelium sustains host-commensal symbiosis in sickness. Nature. 2014;514:638–41. doi:10.1038/nature13823. PMID:25274297.

20. Folseraas T, Melum E, Rausch P, Juran BD, Ellinghaus E, Shiryaev A, Laerdahl JK, Ellinghaus D, Schramm C, Weismüller TJ. Extended analysis of a genome-wide association study in primary sclerosing cholangitis detects multiple novel risk loci. Journal of hepatology. 2012;57:366–75. doi:10.1016/j.jhep.2012.03.031. PMID: 22521342.

21. Goto Y, Obata T, Kunisawa J, Sato S, Ivanov II, Lamichhane A, Takeyama N, Kamioka M, Sakamoto M, Matsuki T. Innate Lymphoid Cells Regulate Intestinal Epithelial Cell Glycosylation. Science (New York, NY). 2014;345:1254009. doi:10.1126/science.1254009.

22. Rausch P, Kunzel S, Suwandi A, Grassl GA, Rosenstiel P, Baines JF. Multigenerational Influences of the Fut2 Gene on the Dynamics of the Gut Microbiota in Mice. Frontiers in microbiology. 2017;8:991. doi:10.3389/fmicb.2017.00991. PMID:28642740.

23. Kashyap PC, Marcobal A, Ursell LK, Smits SA, Sonnenburg ED, Costello EK, Higginbottom SK, Domino SE, Holmes SP, Relman DA. Genetically dictated change in host mucus carbohydrate landscape exerts a diet-dependent effect on the gut microbiota. Proceedings of the National Academy of Sciences of the United States of America. 2013;110:17059–64. doi:10.1073/pnas.1306070110. PMID:24062455.

24. Huse SM, Ye Y, Zhou Y, Fodor AA. A core human microbiome as viewed through 16S rRNA sequence clusters. PloS one. 2012;7:e34242. doi:10.1371/journal.pone.0034242. PMID:22719824.

25. Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R, Mills DA, Caporaso JG. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nature methods. 2013;10:57–9. doi:10.1038/nmeth.2276. PMID:23202435.

26. Davenport ER, Goodrich JK, Bell JT, Spector TD, Ley RE, Clark AG. ABO antigen and secretor statuses are not associated with gut microbiota composition in 1,500 twins. BMC genomics. 2016;17:941. doi:10.1186/s12864-016-3290-1. PMID:27871240.

27. Wang J, Thingholm LB, Skieceviciene J, Rausch P, Kummen M, Hov JR, Degenhardt F, Heinsen FA, Rühlemann MC, Szymczak S. Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota. Nature genetics. 2016;48:1396–406. doi:10.1038/ng.3695. PMID:27723756.

28. Bonder MJ, Kurilshikov A, Tigchelaar EF, Mujagic Z, Imhann F, Vila AV, et al. The effect of host genetics on the gut microbiome. Nature genetics. 2016;48:1407–12. doi:10.1038/ng.3663. PMID:27694959.

29. Langendijk PS, Schut F, Jansen GJ, Raangs GC, Kamphuis GR, Wilkinson MH, Welling GW. Quantitative fluorescence in situ hybridization of Bifidobacterium spp. with genus-specific 16S rRNA-targeted probes and its application in fecal samples. Applied and environmental microbiology. 1995;61:3069–75.

30. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. The ISME journal. 2012;6:1621–4. doi:10.1038/ismej.2012.8. PMID:22402401.

31. Group JCHMPDGW. Evaluation of 16S rDNA-based community profiling for human microbiome research. PloS one. 2012;7:e39315. doi:10.1371/journal.pone.0039315. PMID:22720093.

32. Rodriguez-Diaz J, Garcia-Mantrana I, Vila-Vicent S, Gozalbo-Rovira R, Buesa J, Monedero V, Collado MC. Relevance of secretor status genotype and microbiota composition in susceptibility to rotavirus and norovirus infections in humans. Scientific reports. 2017;7:45559. doi:10.1038/srep45559. PMID:28358023.

33. Smith MI, Turpin W, Tyler AD, Silverberg MS, Croitoru K. Microbiome analysis – from technical advances to biological relevance. F1000prime reports. 2014;6:51. doi:10.12703/P6-51. PMID:25184041.

34. Langille MG, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, Clemente JC, Burkepile DE, Vega Thurber RL, Knight R. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nature biotechnology. 2013;31:814–21. doi:10.1038/nbt.2676. PMID:23975157.

35. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, et al. Diversity of the human intestinal microbial flora. Science (New York, NY). 2005;308:1635–8. doi:10.1126/science.1110591.

36. Holmen Larsson JM, Thomsson KA, Rodriguez-Pineiro AM, Karlsson H, Hansson GC. Studies of mucus in mouse stomach, small intestine, and colon. III. Gastrointestinal Muc5ac and Muc2 mucin O-glycan patterns reveal a regio-specific distribution. American journal of physiology Gastrointestinal and liver physiology. 2013;305:G357–63. doi:10.1152/ajpgi.00048.2013. PMID:23832516.

37. Robbe C, Capon C, Coddeville B, Michalski JC. Structural diversity and specific distribution of O-glycans in normal human mucins along the intestinal tract. The Biochemical journal. 2004;384:307–16. doi:10.1042/BJ20040605. PMID:15361072.

38. Hopkins MJ, Sharp R, Macfarlane GT. Variation in human intestinal microbiota with age. Digestive and liver disease: official journal of the Italian Society of Gastroenterology and the Italian Association for the Study of the Liver. 2002;34(Suppl 2):12–8. doi:10.1016/S1590-8658(02)80157-8. PMID:12408433.

39. Kennedy NA, Walker AW, Berry SH, Duncan SH, Farquarson FM, Louis P, Thomson JM, Satsangi J, Flint HJ, Parkhill J. The impact of different DNA extraction kits and laboratories upon the assessment of human gut microbiota composition by 16S rRNA gene sequencing. PloS one. 2014;9:e88982. doi:10.1371/journal.pone.0088982. PMID:24586470.

40. Lozupone CA, Stombaugh J, Gonzalez A, Ackermann G, Wendel D, Vazquez-Baeza Y, et al. Meta-analyses of studies of the human microbiota. Genome research. 2013;23:1704–14. doi:10.1101/gr.151803.112. PMID:23861384.

41. Kevans D, Turpin W, Madsen K, Meddings J, Shestopaloff K, Xu W, Moreno-Hagelsieb G, Griffiths A, Silverberg MS, Paterson A. Determinants of intestinal permeability in healthy first-degree relatives of individuals with Crohn's disease. Inflammatory bowel diseases. 2015;21:879–87. doi:10.1097/MIB.0000000000000323. PMID:25734694.

42. Ferrer-Admetlla A, Sikora M, Laayouni H, Esteve A, Roubinet F, Blancher A, Calafell F, Bertranpetit J, Casals F. A natural history of FUT2 polymorphism in humans. Molecular biology and evolution. 2009;26:1993–2003. doi:10.1093/molbev/msp108. PMID:19487333.

43. The International Hap. Map Project. Nature. 2003;426:789–96. PMID:14685227.

44. Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD. PANDAseq: paired-end assembler for illumina sequences. BMC bioinformatics. 2012;13:31. doi:10.1186/1471-2105-13-31. PMID:22333067.

45. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI. QIIME allows analysis of high-throughput community sequencing data. Nature methods. 2010;7:335–6. doi:10.1038/nmeth.f.303. PMID:20383131.

46. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics (Oxford, England). 2010;26:2460–1. doi:10.1093/bioinformatics/btq461. PMID:20709691.

47. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Applied and environmental microbiology. 2006;72:5069–72. doi:10.1128/AEM.03006-05. PMID:16820507.

48. Whittaker RH. Evolution and Measurement of Species Diversity. Taxon. 1972;21:213–51. doi:10.2307/1218190.

49. Gauderman WJ. Sample size requirements for matched case-control studies of gene-environment interaction. Statistics in medicine. 2002;21:35–50. doi:10.1002/sim.973. PMID:11782049.

50. Tregouet DA, Ducimetiere P, Tiret L. Testing association between candidate-gene markers and phenotype in related individuals, by use of estimating equations. American journal of human genetics. 1997;61:189–99. doi:10.1086/513895. PMID:9246000.

51. Lozupone C, Lladser M, Knights D, Stombaugh J, Knight R. UniFrac: an effective distance metric for microbial community comparison. The ISME journal. 2011;5:169–72. doi:10.1038/ismej.2010.133. PMID:20827291.

52. Fierer N, Lauber CL, Zhou N, McDonald D, Costello EK, Knight R. Forensic identification using skin bacterial communities. Proceedings of the National Academy of Sciences of the United States of America. 2010;107:6477–81. doi:10.1073/pnas.1000162107. PMID:20231444.

53. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics (Oxford, England). 2004;20:289–90. doi:10.1093/bioinformatics/btg412. PMID:14734327.

54. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic acids research. 2012;40:D109–14. doi:10.1093/nar/gkr988. PMID:22080510.

55. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. Science (New York, NY). 1997;278:631–7. doi:10.1126/science.278.5338.631.

56. Gower JC. Generalized procrustes analysis. Psychometrika. 1975;40:33–51. doi:10.1007/BF02291478.