Gut Origins of Latino Diabetes (GOLD) study

An ancillary study to the NIH Hispanic Community Health Study- Study of Latinos

Multiple Principal Investigators:
Robert C Kaplan, PhD (Contact PI)
Robert Burk, MD

Study protocol
Version 2.0

Table of contents

1. Rationale and specific aims

Hispanics/Latinos, who are the fasting growing group in the US, have a 66% higher prevalence of diabetes compared to non-Hispanic whites (11.8% *vs* 7.1%). Our Hispanic Community Health Study / Study of Latinos (HCHS-SOL) showed diabetes in ~17% and pre-diabetes in 36% of 18 to 74 year old Latino adults. Known diabetes risk factors including diet, physical inactivity, and obesity do not fully explain variation in diabetes across Latinos and other race/ethnic groups. Recent discoveries suggest that the gut microbiome (GMB) may play a role in the etiology of diabetes. GMBs that share similar metagenomic functional groups (e.g., lower butyrate production) tend to be present in diabetic patients, and experimental studies show that transfer of GMB from one individual to another may change the recipient's metabolic health. This raises the possibility that the elevated diabetes risk among Latinos may be partially explained by the influence of the GMB, which in turn might suggest a novel explanation for family and ethnic group clustering of diabetes. However, there are still major knowledge gaps. People living in different geographic locations are known to harbor different potentially adverse sets of GMBs. However, GMB and diabetes has never been studied in a diverse Latino cohort, which is necessary to examine the possibility that population-specific GMB signatures among Latinos may predispose to diabetes. Most prior studies of the GMB and diabetes have used convenience populations and have lacked longitudinal assessments. Studies that overcome these limitations can potentially lead to a new era of diabetes prevention and treatment, since available interventions can modify the GMB.

Our main hypothesis is that specific patterns of the gut microbiome will be associated with diabetes, pre-diabetes and glycemic traits among Latino adults. In HCHS-SOL, the largest ever long-term study of US Latinos, 16,415 cohort members were recruited during 2008-11 **(V1)** and will complete six-year followup examinations during 2014-17 **(V2)**. Both **V1** and **V2** included fasting plasma glucose (FPG) and a 2h oral glucose tolerance test (OGTT), in addition to other laboratory (e.g., HbA1c, insulin), clinical (e.g., adiposity) and behavioral (e.g., diet, exercise) measures relevant to diabetes. With the exception of GMB (fecal) sampling, all of the key protocols are already in place.

Operational tasks are: (i) collect stool samples from 2,000 subjects attending **V2** at all four HCHS-SOL sites, along with clinical and behavioral data including medication use and diet; (ii) analyze stool samples for GMB; and (iii) perform analysis of GMB integrated with extensive existing data in HCHS-SOL (1, 2). Designed in the setting of a large, population based sample, HCHS-SOL is uniquely suited to discern ethnic differences in GMB makeup and to study the association of GMB with diabetes across birthplace/national background groups.

Aim 1. Investigate factors affecting the gut microbiome (GMB) among Latinos. We hypothesize that GMB composition differs with national background (e.g., Mexican, Puerto Rican, Cuban, etc), birthplace (80% are foreign born), gender, age, adiposity, shared household, genetics and relatedness. Diet will also be assessed concurrent with GMB (to update HCHS-SOL V1 diet data) and we will correlate diet with GMB.

Aim 2. Evaluate the association of the gut microbiome (GMB) with the presence of diabetes and pre-diabetes. Our hypothesis is that the GMB differs between across three groups defined at **V2**, including: ***diabetes mellitus*** (N=400, measured FPG≥126; HbA1c≥6.5%; 2hPG≥200 and/or diabetes medication use); ***prediabetes*** (N=800, FPG 100-125; HbA1c 5.7%-6.4%; 2hPG 140–199), and ***normoglycemic*** (N=800, normal FPG, HbA1c and 2hPG). Lab testing from **V2** as well as **V1** (six years earlier) will define groups who meet glycemic criteria persistently over time, and also distinguish recent onset versus longstanding diabetes. Events follow-up among all 1,600 nondiabetics in the GMB study is funded as part of the main SOL study through 2019 and can also identify incident DM.

To help interpret results, we will make use of existing HCHS-SOL data already in place including metabolomics, sociodemographics, health behaviors and genetics (2.5M SNP GWAS array and whole genome sequencing). This will control for important confounders, while integrated analysis of GMB, metabolites and genetic polymorphism data can provide functional insights and point the way to future experimentation and intervention.

2. Background

Epidemic of Diabetes Mellitus in the US Latino Population. Diabetes mellitus (DM) affects over 25 million individuals in the US, with high morbidity and cost (3, 4). Over 50 million people have prediabetes (3). Diabetes does not affect all racial/ethnic groups equally. In a 2007-2009 US survey, 11.8% of Latino adults had diabetes (3), a 66% higher prevalence compared to non-Hispanic whites (7.1%). We recently reported a 17% prevalence of diabetes mellitus in a population-based 2008-2011 sample of Latinos in the US (5). The Latino population comprises the largest and fastest growing US minority population (17% of the US) (6). From 2005 to 2050, the Latino population is projected to account for 60% of the US population growth (6) and will grow to 128 million, or 29% of the entire US (6). Latinos represent diverse cultures, backgrounds, and origins, and as a group tend to experience greater socioeconomic adversity than others. Nationwide, 65% of Latinos are Mexican, 17% Central/South American, 9% Puerto Rican, 4% Cuban, and 5% other background (7). About half of Latino adults are born outside the US (80% of HCHS-SOL are born outside the 50 states), 38% lack a high school education (vs. 9% of non-Hispanic whites) and 25% live in poverty (vs. 11% of non-Hispanic whites). Diabetes is a major health problem for the expanding Latino population and more research is needed to identify new approaches for prevention and management of DM in this and other populations.

Etiology of Type 2 Diabetes The etiology of type 2 DM is multifactorial and has a strong heritable component (20%-80%) based on family, twin, and population studies (for review see (8)). The largest genome-wide association study (GWAS) included 35k cases and 115k controls with individuals of predominantly European ancestry (9). This study identified 63 autosomal loci that together accounted for only 5.7% of variance in disease susceptibility. Additional recent work with multiethnic samples brings the total number of DM related SNPs to 70 (10). Thus, although the GWAS results are promising for particular candidate genes that might explain DM pathogenesis, to date the genetic variants do not explain much of the family-related or heritability of DM. Based on very recent results (11-14), we speculate that the gut microbiome (GMB) will play an important role in diabetes pathogenesis, particularly since the GMB can be acquired from family members (15) and thus, has the potential to explain some family clustering of diabetes (11, 13, 16-18).

Gut Microbiome and Diabetes The human microbiome is the totality of microbes that constitute the community of commensal, symbiotic and pathogenic organisms residing in and on us (19-21). Data supporting the GMB as a cause of diabetes would usher in a new paradigm of disease prevention and management (22-24). The convergence of massively parallel (Next-Gen) sequencing, expanded computing capacities and software pipelines now allows characterization of the microbiome in a culture independent manner empowering new opportunities for research (21, 25, 26). Since the introduction of these transformative technologies, a rapid accumulation of knowledge on the role of the GMB on diabetes and/or metabolic functions is beginning to emerge (for reviews see (12, 16, 20, 21, 27-33)). While it is unclear whether behavioral risk factors for DM may have differential impact across populations, there is convincing data suggesting that people from different global regions harbor different microbiota and the disease associated sets of GMB may be different.

Human studies – Multiple studies (Denmark (17), Sweden (11) and China (13, 18)), including in total < 700 individuals have reported on the association of the GMB and diabetes as reviewed in Nature (12, 14). The most consistent finding amongst the 2 studies that evaluated functional gene groups was the presence of bacterial groups involved in lower butyrate production (12). However, there were striking differences between the 2 populations. Among diabetics, the Swedish study found fewer Clostridiales bacteria (11), whereas the Chinese group found fewer *Roseburia intestinalis* and *F. prauznitzii* organisms (13). Thus, similar functional groups were noted, but they were associated with different microbial organisms in each population demonstrating the need for individualized study in specific populations. In addition, when patients with impaired glucose tolerance (i.e., prediabetes) were characterized (11), > 60% had microbiome features of the group with DM and these individuals also had elevated triglycerides and C-peptide indicating the GMB identified prediabetics with other features similar to those with DM. The model proposed for these findings suggests a scenario where reduction of butyrate producing bacteria leads to intestinal inflammation that has been suggested to be a characteristic of insulin resistance and DM (34). However, it should be noted that both of the studies using a metagenomic approach were performed on individuals with BMIs in the normal range. Thus, there are limited data on individuals that are obese or have excessive BMI as a precursor and risk factor to

development of diabetes. The currently proposed study will have sufficient numbers of high and normal BMI individuals to study the role of the GMB on diabetes development in different strata of BMI.

Animal models – To test the etiological relationship and approach the mechanisms of the GMB on metabolic disorders and diabetes, studies have used gnotobiotic mice (i.e., mice raised in a sterile environment) to transplant GMB bacteria from metabolic abnormal/obese and normal humans into such mice, feed them the same diet and measure weight gain and metabolic function (35). These studies consistently report that the mice receiving the GMB from metabolic abnormal/obese humans gain more weight than the genetically identical mice receiving GMB from lean individuals. The mice also had metabolic abnormalities similar to those seen in human studies (12) and there was an inverse relationship between adiposity and butyrate (35). Other studies modulated the gut microbiome in type 2 diabetic mice models with antibiotics and showed amelioration of insulin resistance with accompanying decrease of inflammatory markers (36-38). Thus, there is animal model support for the relationship between the GMB and metabolic abnormalities/DM.

GMB and diet. Our gut environment houses a complex community of microorganisms (39, 40), which metabolize dietary substrates generating secondary metabolites that influence the host's cardio-metabolic health; in turn, the gut microbiota is itself shaped by the host's diet (41-43). Plant-based diets differ from animal-based diets with respect to several microbe-dependent metabolic pathways, including increased metabolism of fiber and polyphenols, and decreased metabolism of bile acids, choline and L-carnitine, and amino acids (41, 43, 44), which could explain their association with cardio-metabolic risk. For instance, several studies have found a continuous gradient of relative abundances of gut microbial taxa, ranging from high *Prevotella* abundance to high *Bacteroides* abundance; the former is associated with intake of fiber rich plant foods, the latter with animal protein and saturated fat intake (15, 40, 45-48). *Prevotella* species and other microbial strains degrade otherwise indigestible carbohydrates to short chain fatty acids (45), of which butyrate has cardio-protective effects (49). Interestingly, a high *Prevotella*-to-*Bacteroides* ratio has been found to mediate the beneficial effects of dietary fiber on glucose metabolism (50).Taken together, different diets may cause variation in relative abundance of specific gut microbiota , and thus the typical dietary pattern of US Hispanics/Latinos is one of the features of this population that makes the study of the gut microbiome important and interesting.

Impact: Establishing an Etiological Role for the Microbiome in the Pathogenesis of Diabetes The accumulation of consistent evidence for the role of the GMB in diabetes would provide the basis for a major paradigm change in understanding the pathogenesis of diabetes as also discussed by Fredricks and Relman (51). The data generated in this study should provide the needed scientific basis to move to the next stage of the paradigm change - that is in the prevention and treatment of diabetes through targeting the GMB by use of antibiotics, alteration/replacement through probiotics/transplant and diet changes (22-24, 28, 52-54). Current work suggests the gut microbiome is relatively stable over long periods of time (55-59), and making changes to the GMB may require a multimodality approach, e.g., inducing changes through antibiotics or replacement, followed by maintenance of a changed GMB through diet and other means. With extensive host genetic data (GWAS), metabolomics, and dietary data in HCHS-SOL, we are also positioned to examine how cometabolism between microbes and the host may explain the association of GMB with diabetes/prediabetes risk. This type of integrated analysis has been fruitful in other diseases (e.g., role of TMAO in CVD, see Ref (60)), yet little similar information has been collected for diabetes.

3. Methods

Brief overview of research plan. The GOLD study will request gut microbiome (GMB) samples from participants of the HCHS-SOL study returning for visit 2 (**V2**). Out of the 16,415 HCHS-SOL subjects, we will enroll 2,000 subjects to provide a GMB sample. Well-accepted laboratory measures will define diabetes and pre-diabetes. Data will come from **V1** ("baseline" visits, 2008-11), **V2** (six-year follow up visits, 2014-17) and annual HCHS-SOL follow-up contacts. For all enrolled in GOLD, stool collection kits will be distributed through ongoing study contacts and returned by mail for analysis. Longitudinal data on incident diabetes and prediabetes will thus be available both from repeated blood samples, as well as reports of newly-diagnosed cases of incident diabetes that will be ascertained through the main HCHS-SOL study's yearly follow-up.

Study setting: The Hispanic Community Health Study (HCHS) - Study of Latinos (SOL) Cohort

Subjects: HCHS-SOL is a prospective, population-based cohort study of chronic disease risk factors and morbidity and mortality in US Hispanics/Latinos of diverse backgrounds. During 2008-11, 16,415 men and women, of Cuban, Dominican, Puerto Rican, Mexican, and Central and South American background, ages 18-74, were recruited from randomly selected households in defined geographical areas near the 4 field centers (FCs) in San Diego, CA, Chicago, IL, Bronx, NY, and Miami, FL. The cohort was assembled using a stratified two-stage area probability sampling design (61) that provided diversity in SES and national origin or family background. HCHS-SOL oversampled ages 45-74 (60% of cohort) to facilitate examination of target outcomes. On average, 1.8 people per household were enrolled and 9,872 households are represented. Of eligible screenees, 42% were enrolled, a very high percentage for a population sample being recruited into a long-term cohort study. See detailed information on the HCHS-SOL study design, refs (61, 62).

Measures: The baseline assessment, Visit 1 (**V1**), included an examination and collection of laboratory data. See ref (62) and the study web site (www.cscc.unc.edu/hchs/) for full protocols. At **V1**, informed consent was followed by a 7-8 hour in-person baseline examination that included comprehensive biological measurements (e.g., anthropometric measurements, urine and fasting blood samples with ~20 lab tests, oral glucose tolerance test (OGTT, 75 g glucose load), HbA1c, ankle and brachial blood pressures, electrocardiogram). Interviewer-administered questionnaires collected socio-demographic data (e.g., socioeconomic status/SES, place of birth, migration history), medical history, medications, and data on lifestyle risk factors, social and cultural factors, occupation, disability, and health care use. Immediately following V1, participants completed a one-night home sleep study and 7 day accelerometry to estimate physical activity. Diet assessment protocols include two 24-hour dietary recalls were administered, one during V1, and again via telephone 6 weeks later. A food propensity questionnaire containing 115 food items and 137 individual questions was collected 12 months later at the Year 1 phone contact. A 2.5 million SNP GWAS, imputed to ~10 million SNPs, was performed on 13,175 members of the cohort. Whole genome sequencing (WGS) is underway with >4,000 to have completed WGS as well as plasma metabolomic profiling in the near future through the NHGRI Common Chronic Disease Genomics/CCDG collaboration (Kaplan, Boerwinkle).

V2, which is completed ~6 years after **V1**, repeats key study measurements from **V1**. The current proposal will use data from **V1** and **V2** to characterize diabetes and pre-diabetes by fasting plasma glucose/FPG, insulin, hemoglobin A1c, OGTT (2hr glucose and insulin), medication use and diagnoses relating to diabetes. For instance, we can define normoglycemic individuals as those that have normal values at both the first and second visit, while those with diabetes at **V2** can be classified as either recent-onset diabetics or longstanding diabetics (based upon lab tests from **V1** and date of diabetes diagnosis).

Annual Follow-Up: Annual telephone or in-person follow-up interviews are in progress and funded through 2019 to ascertain mortality, incident diabetes, and other outcomes (e.g., CVD events). Of 16,415 participants in the cohort, annual follow-up contacts have been completed by ~85%. To date, over 700 participants have died, and incident "hospitalized" or "fatal events" number in the several hundreds for myocardial infarctions (MIs), strokes, heart failure (HF), and COPD/asthma exacerbations, all of which are undergoing central

adjudication. Among those who were nondiabetic at baseline/V1, to date we identified over 700 cases of incident DM reported during follow-up contacts as a new diagnosis and treatment.

The *Gut Origins of Latino Diabetes* (GOLD) Ancillary Study: Subjects and Design. In the following sections we describe new data collection and use of key existing data for the GOLD ancillary study on GMB and diabetes.

Gut Microbiome Sample Collection : Between Oct 1, 2014 and Sep 30, 2017, all HCHS-SOL participants are being invited to undergo a follow-up clinic visit (**V2**) that will evaluate at least 80% of surviving cohort members (N~12,800). This examination will be conducted during a 6-month window approximately six years after the participants' baseline (**V1**) examination. **V2** is of 3 hours duration, and includes informed consent, interviews, procedures, and exit interview that will include discussion and recruitment for the GMB ancillary study (aka, **G**ut **O**rigin of **L**atino **D**iabetes study, **GOLD**). As with previous and ongoing ancillary studies, we will seamlessly embed recruitment and data collection for **GOLD** into the ongoing V2 examinations.

Retention and Outreach: HCHS-SOL's retention process involves frequent interactions with participants. All receive a quarterly newsletter with articles about health topics of interest to the community, and community events are held relating to the study. Information on the microbiome and health will be included in newsletters and other communications to promote participation in the GOLD ancillary study. FCs emphasize the confidential nature of study data, and provide clinical referrals for problems identified (e.g., diabetes) as well as general assistance (insurance enrollment advice) which are benefits that participants value.

***Logistics of GMB Studies*:**

1. The Einstein team (Kaplan, MPI) will coordinate study set up including finalizing protocols, IRB applications, materials to advertise the study among HCHS-SOL participants, staff training, and interaction among the sites.
2. All 4 HCHS-SOL field centers will collect fecal specimens and ship to MPI Burk's lab (specimen repository).
3. After intake and tracking of specimens, Dr. Burk will ship specimens in batch to MPI Knight's lab (GMB lab).
4. The Knight lab will be responsible for extraction, amplification, next-gen sequencing and the first level GMB analysis including filtering, normalization, creating OTU tables and functional gene groups from sequence data.
5. GMB data from Dr. Knight will be provided to the statistical team at Einstein for epidemiological analyses. Dr. Knight's lab will also be involved in the ecological analyses and visualization of data (21, 63-65).
6. The HCHS-SOL Coordinating Center will create on line systems for the sites and laboratories to use in order to track eligibility, enrollments, refusals, and achievement of enrollment targets.

Recruitment approach. To date, HCHS-SOL is on track to meet its goal of recruiting 12,800 participants from the initial cohort (80%) to return for the 2nd visit (V2). Each of the 4 sites will request that the participants attending V2 participate in GOLD to help understand the role of the GMB on their health. Individuals that agree to participate will provide written informed consent. They will be instructed in use of a specimen collection kit that will be sent home with them containing all the materials needed. A pre-addressed stamped envelope will allow easy return of the specimen to the Burk lab. A brief in person questionnaire will elicit use of antibiotics and GI medications within 6 months of sample collection so that we can assess the impact on GMB and perform appropriate exclusions during analyses (21). Participants that return a specimen will receive modest compensation. GMB (fecal) specimen will be collected at a single V2 time point, which will be appropriate based on the observations that a GMB sample is representative of an individual's core GMB and is stable for years if not decades (55, 59, 66).

Upon recruiting a participant into the GOLD ancillary study, field center staff will complete the "GOLD Enrollment and Tracking Form" (GOL) to document the fact and date of enrollment. A brief questionnaire (GLQ) will record information about medication use, diet and medical history. Both of these forms will be completed at the time of recruitment, after which time the participant will be sent home with the specimen collection kit.

Methods for microbiome sampling. Gut microbiome (fecal samples) will be self-sampled using a disposable paper inverted hat (Protocult collection device, ABC Medical Enterprises, Inc., Rochester, MN) that goes over the toilet seat. Two samples will be obtained using a method that will be provided and explained by study staff, with an instructional video providing information about how to collect the stool. The participants will sample the stool with a plastic applicator, spreading a small amount on a Whatman FTA card (Sigma-Aldrich). They will collect a second sample which will be placed in a supplied container containing a stabilizer (RNAlater, Qiagen, Valencia, CA, OmniGene Bioproducts, Woburn, MA), shaking the tube in order to mix the stool and the preservative which stabilizes DNA and RNA (67). Specimens collected in this manner have been shown to preserve the composition of the microbiome for up to 14d at room temperature (68). Both specimens are placed into a bag with absorbent / desiccator material and sent 'Priority Mail Express' (USPS) in a supplied envelope to the Burk lab. Thousands of samples have been collected using mail-in approaches by the American Gut Project (http://americangut.org), run by our MPI, Dr. Rob Knight. Kits will be prepared by the Burk laboratory for distribution to all sites. When the fecal samples are received by Dr. Burk, they will be assigned a lab ID, logged into the database and stored at -80^{o} C until processing for molecular analyses. The lab ID is linked to the kit ID. The kit ID can be unequivocally linked to each subject. No subject's personal identifying information (e.g., name) need be included on the kit.

Data Capture and Management: A centralized web-based data entry system was developed and implemented by the Coordinating Center (CC) for **V1** and **V2**. The system developed by the CC has several innovative components that include a simple, menu-based interface that allows detecting local data problems, and report generation features for data completeness and quality reports. This allows immediate detection and correction of data errors at the FCs. The CC will add data management modules to support the GOLD study.

The Burk laboratory, which will receive all specimens by mail from GOLD participants at the four field centers, will provide each field center with an updated listing of specimens received. Regular conference calls and correspondence will address issues with specimen quality, completeness and other logistical problems. Upon receiving confirmation from the Burk laboratory that the stool specimen has been received, field centers will provide monetary reimbursement to the GOLD study participant according to local field center procedures.

***Classification of GOLD Study Outcomes: Diabetes and Prediabetes*:** Key study procedures include fasting blood draw and a 2-h oral glucose tolerance test (OGTT for those free of diabetes). Participants are required to fast for at least 8-h prior to the visit consuming only water and necessary medications. Venous blood specimens are collected, processed and frozen on-site towards the beginning of the visit and also 2-h after a 75g glucose load and then analyzed in batches. Plasma glucose is assessed using a hexokinase enzymatic method (Roche Diagnostics Corporation, Indianapolis, IN). Glycosylated hemoglobin (HbA1c) is measured in EDTA whole blood using a Tosoh G7 Automated HPLC Analyzer (Tosoh Bioscience Inc., SF, CA). Self-reported information is used to define personal and family history of diagnoses including diabetes. Inventory methods are used to list all currently taken medications. Definition of diabetes and prediabetes are based on standard ADA criteria of laboratory tests (69).

Diabetes is defined as FPG ≥ 126 mg/dL (7 mmol/L); a 2-h post-load glucose level (2hPG) ≥ 200 mg/dL (11.2 mmol/L); A1c level ≥ 6.5%; and/or self-reported and/or documented use of antihyperglycemic drugs based upon medication labels scanned at study visits or reported by participants.

Pre-diabetes is defined as FPG ≥ 100 mg/dL (5.5 mmol/L) AND < 126 mg/dL (7 mmol/L); a 2-h post-load glucose level (2hPG) ≥ 140 mg/dL (7.7 mmol/L) AND < 200 mg/dL (11.2 mmol/L); A1c level ≥ 5.7% (39 mmol/mol) AND < 6.5%. No reported and/or documented use of antihyperglycemic medication.

Normoglycemic individuals are defined on the basis of FPG < 100 mg/dL (5.5 mmol/L); a 2-h post-load glucose level (2hPG) < 140 mg/dL (7.7 mmol/L); A1c level < 5.7%. No reported and/or documented use of antihyperglycemic medication. To define a true normal control population, we will require that individuals have normal levels of all glycemic variables and no diabetes medications at **V1** and **V2**.

Worsening glycemia will be defined by FPG and A1c level across a 30 month interval in the longitudinal part of the study.

Incident diagnosed diabetes, among individuals who were pre-diabetic or normoglycemic at **V2**, is defined as self-report of diabetes during follow-up after **V2** as ascertained during annual followup contacts.

Molecular Characterization of the Gut Microbiome

Laboratory methods: The fecal microbiome samples will be processed for DNA from the stool samples as described by the Knight lab using a modification of the MoBio (Carlsbad, CA) 96-well PowerSoil DNA isolation kit that employs both chemical and physical (i.e., bead beating) means to release DNA for precipitation and amplification (21). We will use the V4 primer pairs (515F/806R) that amplify a 322 bp fragment (70) that will undergo sequencing with paired-end reads on the HiSeq or MiSeq Illumina NGS platforms. These primers and methods have been shown to have the highest sensitivity for capturing bacterial diversity (55). To generate a full length V4 16S rRNA amplicon sequence, the paired-end reads will be joined into a single sequence (using the read overlaps) with the FLASH algorithm (version 1.2.11) (71).

Microbiome sequence analyses: The computational analyses will include recently introduced quality-filtering methods that are critical for accurate taxonomic assignment of individual reads (72). The core microbial processing pipeline will be QIIME (Quantitative Insights Into Microbial Ecology) (2, 64). It combines several programs for the analysis of 16S rRNA gene amplicon data. The process begins with taxonomy-independent binning of the sequences according to their degree of sequence similarity. We will bin the reads according to different criteria of similarity, from > 97% (usually defined as species level) to lower taxonomic resolution. The clusters will undergo annotation - using BLAST against the available curated 16S amplicon databases such as Greengenes (73, 74) and SILVA (75). In parallel, phylogeny is calculated using the pyNAST algorithm to create a phylogenetic tree (76). Recently, for improved classification of V4 or other 16S reads, the Knight Lab has been compiling a database that includes many large datasets, e.g., the Human Microbiome Project (77, 78) and Earth Microbiome Project (79) datasets, all beginning with the raw data and processing the reads through the same systematic pipeline. This is especially important because technical decisions in a pipeline can have larger effects on the scientific interpretation, (e.g., the ratios of phyla such as Firmicutes, Bacteroidetes, and Actinobacteria that have been previously linked to metabolic abnormalities (80-82), than do the underlying sequences (83). This database, available at http://www.microbio.me/qiime, at the time of this writing, contains 143,965 human and environmental samples from 835 studies.

QIIME can also provide insight into a range of key microbial community parameters, for example: **(i) alpha diversity** (i.e., how many kinds of microbes are in a given community, or how much branch length it covers in a phylogenetic tree), **(ii) beta diversity** (how microbes partition across different communities), including the UniFrac measure introduced by the Knight lab to perform calculations of community difference in a phylogenetically informed way (84, 85), and **(iii) PICRUSt**, which stands for Phylogenetic Investigation of Communities by Reconstruction of Unobserved States, is a method for leveraging phylogenetic information such as that obtained inexpensively by 16S rRNA surveys into **functional information/groups** (1). Essentially, known genomes are converted into vectors of molecular functions (e.g., Kegg Orthology terms), and interior nodes are inferred by ancestral state reconstruction (86). New lineages for which no known genome is available are placed into the tree by any one of several insertion methods, and the maximum likelihood estimate of the function vectors (of these vectors) is the ancestral state of the node to which they connect in the tree. Importantly, the more genomes are available from a given environment, the better PICRUSt works, and we can obtain not just an estimate of the functions in each genome but also an estimate of how accurate those predictions are by comparison to the nearest neighbors. For human-associated environments, correlations are generally in the r=0.8 range, and in the human gut, where reference genomes are most dense, correlations are frequently in the r=0.9 range. This is despite processes such as horizontal gene transfer, which are especially prevalent in the human gut (87); the explanation for this is that different gene transfers happen with different genes and between different pairs of lineages, so no clear alternative phylogenetic signal distracts from the overall tree (88). Since its publication in 2013, PICRUSt has been used in over 50 published studies. For example, David et al., 2014 (89) used it to infer microbial community functional changes that occur rapidly with diet, showing that even within a subject over time the functional changes can be estimated accurately from the phylogenetic profile (published in Nature). Similarly, Dassi et al., 2014 (90) were able to use PICRUSt to understand functional changes in the oral microbiota associated with

probiotics, and Davenport et al., 2014 (100) were able to identify seasonal fluctuations in function in the human gut microbiome. PICRUSt is thus a well-validated and increasingly widely used tool to extrapolate functional gene groups from 16S rRNA taxonomic profiles at a fraction of the cost of shotgun metagenomics.

We will also perform shotgun metagenomics directly on as many of the samples as possible, using the HUMANn2 pipeline, an elaboration of the HUMANn pipeline (91) developed for the Human Microbiome Project (39), to extract taxonomic profiles and functional inventories from the shotgun metagenomic data, and compare these profiles to the PICRUSt profiles to re-validate the pipeline for our population. We will further analyze the shotgun metagenomic data using ConStrains (39), an algorithm we recently introduced to perform strain-level profiling without the need for multiple reference genomes in each taxon, perform metagenome assemblies using MetaSpades [http://bioinf.spbau.ru/en/spades3.7], developed by our close collaborators in the Pevzner lab at UCSD, and perform horizontal gene transfer analysis using DarkHorse (92), also developed by our close collaborators at UCSD. These functional and taxonomic tables will be converted into BIOM-format tables (93) and exported to QIIME (2) for downstream analysis analogous to that for 16S rRNA taxonomy tables, including ANCOM (94) for differential abundance analysis including explicit handling of the compositionally of the data, and exported to R using phyloseq (95) for additional biostatistical analysis.

Statistical Analysis approaches

The microbiome data from each subject will be available for biostatistical analyses to address the specific aims in a number of different formats from the Knight Lab. The taxonomic/functional tables annotation and phylogenetic information for each sample will constitute a subset of parameters that will serve as input for statistical analysis. In addition, the inferred functional gene families will be available for each sample as a metagenome table with abundance of gene families (i.e., functional gene groups) (1). For instance, functional gene clusters will include sets that are involved in carbohydrate and lipid metabolism, e.g., fatty acids, carbohydrates, etc. (1). All data analyses will be preceded by extensive data checking and verification to identify and resolve the reasons for missing values, inconsistencies, and out-of-range values. Weights that account for sampling probability, nonresponse and household clustering in HCHS-SOL are derived and accounted for in analyses.

In Specific Aim 1, we will evaluate the association of demographic and other factors with gut microbiome profiles among those with normal indices of glucose metabolism (normal fasting and 2-hour OGTT glucose levels, HbA1c < 5.7). Candidate predictors of the microbiome will include gender; age; place of birth; national background; BMI; smoking; renal & liver function; lipids; inflammation; medications (e.g., antibiotic use). We will also investigate dietary patterns including fiber, meat, fruit/vegetable, fats, energy intake, & alcohol intake as predictors of GMB. For missing variables, we will first examine the missingness mechanism by checking if missingness is related to other variables. If missingness is considered to be ignorable (missing at random (MAR) or missing completely at random (MCAR)), multiple imputation will be performed (96, 97). Additionally, sensitivity analyses will evaluate the robustness of imputation.

We will first explore the microbiome data by unsupervised hierarchical clustering analysis. Specifically, we will calculate dissimilarity (e.g.the uniFrac distance, Jensen-Shannon distance) between each pair of microbiome profiles and use the dissimilarity matrix as the input for cluster analysis (e.g PAM clustering). We expect microbiome profile of individuals with similar characteristics to be clustered together. We will evaluate which characteristics are associated with cluster assignment by χ^2 tests. To assess the contribution of household to the microbiome, the dissimilarity will be computed between pairs of first degree relatives, pairs of second degree relatives, pairs of unrelated individuals from the same household, and pairs of randomly selected unrelated subjects from different households. The genetic relationship of subjects in HCHS-SOL (kinship) has been derived by the software KING based on GWAS data(98). In total, there are 10,861 mutually unrelated individuals, and well over 2,000 closely related individuals which includes ~1,000 parent-offspring duos and 211 trios (two parents and offspring). Shifts in the distributions of the dissimilarity across the specific groups of interest will then be evaluated with the Wilcoxon test. We will also perform principal coordinate analysis, hierarchical clustering and other methods (99) to identify communities of microbes (e.g., enterotypes) and analyze functional enrichment of distinct groups (1, 48). This will be followed by supervised analyses of specific microbiome taxa (and/or functional gene group). Specifically, the association of a specific binary or

categorical variable (e.g., gender; geographic background) with a specific taxon (and/or functional gene group) will be assessed by comparing the mean abundance between groups with the nonparametric T-test for two group comparisons and MANOVA for more than two groups in which p-values will be evaluated by permutation. The Spearman correlation and/or association metrics (100) will be used to evaluate the association of abundance of different taxa (and/or functional gene group) with continuous variables. To adjust for the false-discovery rate due to performing multiple testing of the individual microbiome taxa (and/or functional gene group), we will apply the Benjamini-Hochberg procedure (101).

We will then perform joint analyses for simultaneously modeling the relationship of multiple clinical and laboratory variables and GMB profile. These multivariable analyses assess the independent effect of each predictor (e.g., age, gender, BMI) while adjusting for the effects of the others. To do this, we will use Dirichlet-multinomial regression (102). Specifically, let $z_i = (z_{i1}, ..., z_{iq})$ be the normalized read counts in *q* microbial features (e.g. OTUs) for individual i and $x_i = (x_{i1}, ..., x_{ip})$ be the vector of *p* covariates. The Dirichlet-multinomial regression models the read counts using the multinomial distribution with parameter $\pi_j(j = 1, ..., q)$, where π_j is the probability for the *j*th feature. The model then assumes $\pi_j(j = 1, ..., q)$ follows a Dirichlet distribution with parameters $\gamma_j(j = 1, ..., q)$. To link the covariate effects and $\mathrm{microbial}$ counts, γ_j is modeled as a function of the covariate vector: $\gamma_j(x_i; \beta_{j*}) = exp(\beta_{j0} + \sum_{k=1}^{p} \beta_{jk} x_{ik})$. If a covariate *k* has no effect on the microbiome profile, then $\beta_{1k} = ... = \beta_{qk} = 0$. This null hypothesis can be tested using the likelihood ratio test. To identify the variables associated with specific microbial features, sparse penalties will be applied to the coefficients to select the non-zero β_{jk}, which correspond to the effect of covariate *k* on feature *j*. In multivariate analyses, we will use our detailed knowledge of the cohort to avoid false inferences due to collinearity (e.g., Puerto Ricans are concentrated in New York and are more likely to be US born).

Additional analysis: With genetic markers from the 2.5M SNP genotyping Illumina panel (1000G imputed to 10M SNPs), we have identified clusters of individuals that share common ancestral origins (103-105). We will test if genetic ancestry is associated with microbiome profile, independent of diet and other confounders. As an exploratory analysis, we will use GWAS to identify any SNPs that influence the microbiome composition and/or individual microbial feature. While main analyses will be conducted in the normoglycemic group (n=800), we will repeat analyses with pre-diabetics included to examine the generalizability to this larger group of the identified correlates of GMB (total n=1,600). We will limit analyses to weight stable individuals (V1 versus V2) in secondary analyses to reduce the influence of individuals likely to have had substantial changes in GMB or diet over time.

In Specific Aim 2, we will first characterize the diversity of the microbiome among the diabetes, pre-diabetes and normal groups by comparing the number of observed microbial features using ANOVA (see (15). As noted above, we will use functional gene families (e.g. those assembled from 16S rRNA reads by PICRUSt (1)) and reiterate the analyses. We will test the effect of microbiome composition on disease status using kernel-based regression by treating the disease status as the outcome and microbiome composition as a predictor. The advantages of kernel regression include: (1) it models complex non-linear interactions among microbial features; and (2) the kernel (e.g. based on uniFrac distance) implicitly incorporates phylogenetic information in the model. We will model $g(\mu_i, x_i, z_i) = x_i^T \beta + h(z_i)$, where g(.) is a link function (logit link) and μ_i is the mean of outcome y_i ($y_i = 0, 1, 2$ for the three groups). If microbiome composition has no effect on outcome, then $h(z_i) = 0$. The kernel-based regression models will use the kernel function: $h(\cdot) = \sum_i \alpha_i K(\cdot, z_i)$, where the kernel $K(\cdot, \cdot)$ is the distance metric between two microbiome profiles (106). The kernel-based regression allows for testing and estimating the effects of microbiome composition on disease outcome as well as adjusting for the effects of covariates such as BMI. Once we estimate $h(z_i)$, we can calculate the probability of disease outcome given a new sample of GMB. The model's ROC curve will be validated using k-fold cross-validation.

We will then identify specific microbiome features associated with diabetes/prediabetes status. Our analysis includes two steps: (1) ANOVA followed by pairwise comparisons of the following three groups: normoglycemic, pre-diabetic and diabetic; (2) random forest (RF) models using microbial features found to be significant from the first step. RF is an ensemble machine learning technique using decision trees as base

learners and allows us to predict disease status based on microbiome features. The RF model will be validated incorporating all steps of the analysis by k-fold cross-validation.

Power Considerations: Statistical power for microbiota studies relies on multiple parameters including the relative abundance and inherent variability of each taxon in a diverse community. As such, we estimated the proposed study's power by a simulation study based on the beta-binomial distribution used in White et al. (107), in which sequences were randomly sampled from a population mean proportion and a dispersion value, allowing for different biological variability levels among samples. We considered microbiome taxa with > 100 reads (roughly equivalent to relative abundance ≥1% with 10,000 16S reads per sample) and set the dispersion parameter to be 10e-3. The **Table (at right)** shows the estimated power (%) to detect the association of an epidemiological exposure (**Aim 1**) and disease outcomes (**Aim 2**) as a function of fold changes (FCs) of abundance between two comparing groups. As expected, increased power was found with higher relative abundance. For a low abundance (1%, 100 read counts), Aim 1 will have >90% power at the significance level of 0.00001 (Bonferroni-corrected p-value given our anticipated 1000 independent tests), to detect a 16% change of abundance for a taxon when the exposure rate is at 30%. Moreover, with a higher relative abundance (>5%), Aim 1 will have excellent power (>90%) to detect a 16% change even when the exposure rate is relatively low (0.2 and above). For Aim 2, for a taxon with an abundance of 5%, we will have >90% power to detect 6% fold change between diabetes (n=400) and normal controls (n=800) or between pre-diabetes (n=800) and normal controls (n=800). We will also examine top microbiome taxa associated with outcomes in Aim 2 in relation to new development of diabetes (report of new diagnosis/use of diabetes medications at a subsequent annual follow-up interview) among 1,600 non-diabetic subjects with GMB. We expect a minimum of 96 individuals to develop diabetes during 3 years follow-up (based on an estimated rate of conversion of 2-3% per year in our cohort). With a minimum of 96 cases of diabetes, the study will have >80% power when α=0.05 to detect a hazard ratio of 1.33 or greater for a 1 standard deviation increase in the abundance of a specific GMB taxon.

Aim 1: Predictors of GMB, n=800

Relative abundance	Exposure rate	FC=1.14	FC=1.16
0.01	0.3	0.79	0.94
	0.4	0.89	0.98
0.05	0.3	0.79	1.00
	0.4	0.89	1.00

Aim 2: Diabetes vs Normals, n=400 vs n=800

Relative abundance	FC=1.08	FC=1.10
0.05	>0.99	>0.99

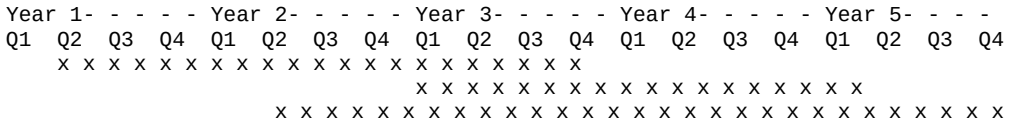
Aim 2: Prediabetes vs Normals, n=800 vs n=800

Relative abundance	FC=1.08	FC=1.10
0.01	0.63	0.93
0.05	>0.99	>0.99

FC, Fold Change

Rigor and transparency HCHS-SOL protocols are all published on a website (www.cscc.unc.edu/hchs/). Publications are reviewed by the study committees before submission to a journal, and all data analysts are subjected to replication of data analysis by the Coordinating Center. All staff are centrally monitored and trained, with ongoing review and QC of data. As an ancillary study, the HCHS-SOL Steering Committee and NHLBI program staff as well as Observational Study Monitoring Board members have approved our proposal.

Timeline



4. Protection of human subjects

1. RISKS TO PARTICIPANTS

1.1. Study setting: HCHS/SOL is an NIH-initiated longitudinal cohort study of risk and protective factors for various conditions including heart disease, type 2 diabetes, and pulmonary disease. A cohort of 16.415 Hispanic adults 18-74 years was recruited during 2008-2011 from four field centers Albert Einstein College of Medicine, Bronx; Northwestern University, Chicago; University of Miami, and San Diego State University). HCHS/SOL investigators are conducting a second in-person examination cycle that started in October 2014 and will continue through 2017. It is estimated that 80% of the original cohort will be re-examined (N = 12,800). After this second examination, the cohort will be followed-up to quantify all-cause mortality, fatal and non-fatal CVD, diabetes, and pulmonary disease. Exclusion criteria at the initial HCHS/SOL enrollment contact included plans to move beyond commuting distance within three years, any cognitive or physical impairment that would interfere with informed consent and completion of study visits, or any other circumstances that might interfere with data collection and follow-up efforts.

1.2. Study participants: In this ancillary study we propose to enroll 2,000 HCHS/SOL participants who are coming to the HCHS/SOL clinic for their second examination, equally distributed across the four field centers. We anticipate that the sampling distribution will follow that of the overall HCHS/SOL sample (e.g. 60% female). We will not impose any additional inclusion criteria beyond those already in place.

1.3. Collaborating sites:

Coordinating Center (CC): the coordinating center will be an administrative unit in the study, with no recruitment or assessment being administered by any of its staff, and will not interact directly with study participants. The only direct risk to study participants resulting from CC activities is a violation of privacy. The CC's primary responsibility is ensuring confidentiality of the study data through the development of rigorous study protocols and maintaining standard operating procedures required to secure the network and study datasets.

Bronx Microbiome Coordinating Center (BMCC) at Albert Einstein College of Medicine: Participants will mail specimens directly to the BMCC for processing, but the BMCC will not interact directly with study subjects. Any risk of violation of privacy is minimized by having specimen collection tubes de-identified with unique codes that the BMCC cannot trace back to an individual participant. The BMCC sends weekly and monthly reports of specimens received to field centers and the coordinating center for monitoring study progress.

Dr. Robert Knight research laboratory (University of California San Diego): This laboratory will conduct the analyses of the gut microbiome from specimens processed by the BMCC. They will not interact directly with study subjects. All samples received will be de-identified.

Field centers: Participants will be enrolled from the four HCHS/SOL field centers (Albert Einstein College of Medicine; University of Miami; San Diego State University; and University of Illinois-Chicago). The protocol imposes minimal risk to subjects providing stool specimens and questionnaire data. These risks are considered minor, and generally do not go beyond risks experienced in routine clinical practice.

1.4. Sources of materials: We will collect stool specimens from study participants. Participants will be asked to provide stool specimens according to a standardized protocol, for which they will receive verbal and video instructions. Participants will be asked to mail their specimens to the BMCC using self-addressed and self-stamped packets provided with the specimen collection kits. The collection kit contains all the materials needed to prepare the samples and mail them (e.g. tubes, disposable gloves, envelopes). In addition, a dataset will be prepared by the CC which will include participants' demographic, medical health and phenotypic information that is available from the HCHS/SOL baseline and second examination.

1.5. Potential risks: Providing stool specimens poses no or minimal risk and the protocol is consistent with what participants may encounter in clinical practice. A risk to participants is the loss of confidentiality due to

breaches in data security. It is possible that a participant might experience embarrassment or discomfort upon collecting their specimens and could experience the protocol as an inconvenience.

2. ADEQUACY OF PROTECTION AGAINST RISKS

2.1 Recruitment and Informed Consent: Recruitment will be conducted by trained and certified research personnel at each HCHS/SOL field center. Participants will be approached after completion of their HCHS/SOL examination, before they leave the clinic. Research staff will introduce the study to participants and obtain written consent if they are willing to participate. After obtaining consent, participants will receive instructions about how to collect specimens, and will be given written instructions and the specimen collection kit. Participants' informed consent will be obtained at each field center before any data collection activity by trained research staff. The informed consent form will be based on a study-wide document, modified if necessary to meet local IRB requirements. The informed consent form will describe the objectives of the study, the voluntary nature of participation, the procedures involved in the study and potential risks associated with participation. Participants will be informed that they have the right to withdraw from the study at any time without affecting their relationship with HCHS/SOL or other ancillary studies or their doctors. Participants will be also asked to state their willingness to have their data shared with scientific collaborators outside the local field center and study investigators. The Coordinating Center will develop an informed consent tracking system that allows for changes of levels of consent provided by participants if they change their mind and request a modification of their original consent. This tracking system will facilitate removal of participants' samples from repositories when necessary.

2.2. Protections against risk

a. Confidentiality and data security: Confidentiality of data is protected through storing all data only with numerical codes. The study computer systems will provide file security by requiring a username and password for access so that confidential data are not released. All systems are located behind University firewalls. Data containing any identifying information (e.g. name, address) will be kept confidential from the study ID. Any printed materials containing information on participants will be stored in locked filing cabinets within locked offices. Data for the study transferred from field centers and laboratories to coordinating center will take place over a secure network with password protection and encryption technology. Study data is stored on a server that is protected by a hardware firewall. User logins and passwords are assigned to authorized personnel only, with IP address restrictions. The system complies with HIPAA requirements.

b. Reporting of results: Because results from gut microbiome analyses do not have clinical relevance yet, we will not report these results back to participants. This would be made clear during the informed consent process. Fasting glucose and hemoglobin A1c measurements to be performed will be returned to participants, who will be advised to seek medical attention should abnormal levels be identified. Participants will be advised that we are not testing for blood in their stool, nor will we use the specimen to test for colorectal cancer, and that participation in the GOLD ancillary study does not substitute for medical screening for diseases.

3. POTENTIAL BENEFITS OF THE PROPOSED RESEARCH TO THE SUBJECTS AND OTHERS

No direct benefits to the study participant are expected to result from participation in the proposed study, as will be stated in the informed consent. Potential benefits to the Hispanic population and the population at large by elucidating the role of gut microbiome on diabetes risk.

4. IMPORTANCE OF KNOWLEDGE TO BE GAINED

Hispanics are the largest minority group in the U.S. Data from national studies and HCHS/SOL indicate that obesity and diabetes are important health problems in this population. Prevention of obesity and diabetes, and their precedent risk factors presents the best means to reduce the morbidity and mortality arising from these conditions. Because therapeutic interventions are possible to target the gut microbiome, understanding the relationship of the gut microbiome to diabetes and other conditions is anticipated to lead to a whole new era of prevention and treatment options.

5. The Investigative Team The team combines complementary skills including epidemiology, genetics, biostatistics, computer science, and molecular ecology. Dr. Robert Burk (Einstein, MPI) is a physician-scientist trained in Medical Genetics with over 30 years of translational / molecular epidemiology research experience. Kaplan (Einstein, corresponding MPI), an epidemiologist, is a leader in diabetes and cardiovascular epidemiology cohort studies and a PI and Chair of the Steering Committee of HCHS-SOL. Dr. Rob Knight (UC-San Diego, MPI) is an internationally recognized expert on the microbiome who will lead the analysis of the fecal samples and bring to bear the latest innovations in GMB analysis. Wang (Einstein) is the project biostatistician who brings understanding of the complex nature of the microbiome data applied to a large cohort study.

As a multicenter study, this ancillary study to HCHS-SOL will be conducted in parallel at three other field centers, using identical procedures. Drs. Talavera (San Diego State U.), Daviglus and Perkins (U. of Illinois-Chicago), Gellman (U. of Miami), Davis (Coordinating Center, UNC) are the PIs at the HCHS-SOL Field Centers providing nation-wide expertise in Latino diabetes and field work. Our collaborator Dr. Kari North (University of North Carolina) from the HCHS-SOL Coordinating Center will be responsible for coordinating the activities of the multiple HCHS-SOL centers.

Contact information for the lead PI and Study coordinator are as follows:

Principal Investigator: Dr. Robert C. Kaplan
Department/Division: Epidemiology and Population Health
Email Address: robert.kaplan@einstein.yu.edu
Phone Number 718-430-4076

Study Contact/Project Coordinator: Madeline Crespo-Figueroa
Study Contact/Project Coordinator Email Address: madeline.crespo-figueroa@einstein.yu.edu
Study Contact/Project Coordinator Phone Number: (718)584-1563

6. References

1. Langille MG, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, Clemente JC, Burkepile DE, Vega Thurber RL, Knight R, Beiko RG, Huttenhower C. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nature biotechnology. 2013;31(9):814-21. doi: 10.1038/nbt.2676. PubMed PMID: 23975157; PubMed Central PMCID: PMC3819121.
2. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7(5):335-6. doi: 10.1038/nmeth.f.303. PubMed PMID: 20383131; PubMed Central PMCID: PMC3156573.
3. Prevention CfDCa. National diabetes fact sheet: national estimates and general information on diabetes and prediabetes in the United States, 2011. USDoHaH Services, Editor; 2011.
4. American Diabetes A. Economic costs of diabetes in the U.S. in 2012. Diabetes care. 2013;36(4):1033-46. doi: 10.2337/dc12-2625. PubMed PMID: 23468086; PubMed Central PMCID: PMC3609540.
5. Daviglus ML, Talavera GA, Aviles-Santa ML, Allison M, Cai J, Criqui MH, Gellman M, Giachello AL, Gouskova N, Kaplan RC, LaVange L, Penedo F, Perreira K, Pirzada A, Schneiderman N, Wassertheil-Smoller S, Sorlie PD, Stamler J. Prevalence of major cardiovascular risk factors and cardiovascular diseases among Hispanic/Latino individuals of diverse backgrounds in the United States. JAMA : the journal of the American Medical Association. 2012;308(17):1775-84. doi: 10.1001/jama.2012.14517. PubMed PMID: 23117778; PubMed Central PMCID: PMC3777250.
6. Passel J, Cohn D. US Population projections 2005–2050. Washington, DC: Pew Research Center, 2008.
7. Motel S, Patten E. The 10 largest Hispanic origin groups: characteristics, rankings, top counties. Washington, D.C.: Pew Hispanic Center, 2012.
8. Ali O. Genetics of type 2 diabetes. World journal of diabetes. 2013;4(4):114-23. doi: 10.4239/wjd.v4.i4.114. PubMed PMID: 23961321; PubMed Central PMCID: PMC3746083.
9. Morris AP, Voight BF, Teslovich TM, Ferreira T, Segre AV, Steinthorsdottir V, Strawbridge RJ, Khan H, Grallert H, Mahajan A, Prokopenko I, Kang HM, Dina C, Esko T, Fraser RM, Kanoni S, Kumar A, Lagou V, Langenberg C, Luan J, Lindgren CM, Muller-Nurasyid M, Pechlivanis S, Rayner NW, Scott LJ, Wiltshire S, Yengo L, Kinnunen L, Rossin EJ, Raychaudhuri S, Johnson AD, Dimas AS, Loos RJ, Vedantam S, Chen H, Florez JC, Fox C, Liu CT, Rybin D, Couper DJ, Kao WH, Li M, Cornelis MC, Kraft P, Sun Q, van Dam RM, Stringham HM, Chines PS, Fischer K, Fontanillas P, Holmen OL, Hunt SE, Jackson AU, Kong A, Lawrence R, Meyer J, Perry JR, Platou CG, Potter S, Rehnberg E, Robertson N, Sivapalaratnam S, Stancakova A, Stirrups K, Thorleifsson G, Tikkanen E, Wood AR, Almgren P, Atalay M, Benediktsson R, Bonnycastle LL, Burtt N, Carey J, Charpentier G, Crenshaw AT, Doney AS, Dorkhan M, Edkins S, Emilsson V, Eury E, Forsen T, Gertow K, Gigante B, Grant GB, Groves CJ, Guiducci C, Herder C, Hreidarsson AB, Hui J, James A, Jonsson A, Rathmann W, Klopp N, Kravic J, Krjutskov K, Langford C, Leander K, Lindholm E, Lobbens S, Mannisto S, Mirza G, Muhleisen TW, Musk B, Parkin M, Rallidis L, Saramies J, Sennblad B, Shah S, Sigurethsson G, Silveira A, Steinbach G, Thorand B, Trakalo J, Veglia F, Wennauer R, Winckler W, Zabaneh D, Campbell H, van Duijn C, Uitterlinden AG, Hofman A, Sijbrands E, Abecasis GR, Owen KR, Zeggini E, Trip MD, Forouhi NG, Syvanen AC, Eriksson JG, Peltonen L, Nothen MM, Balkau B, Palmer CN, Lyssenko V, Tuomi T, Isomaa B, Hunter DJ, Qi L, Wellcome Trust Case Control C, Meta-Analyses of G, Insulin-related traits Consortium I, Genetic Investigation of ATC, Asian Genetic Epidemiology Network-Type 2 Diabetes C, South Asian Type 2 Diabetes C, Shuldiner AR, Roden M, Barroso I, Wilsgaard T, Beilby J, Hovingh K, Price JF, Wilson JF, Rauramaa R, Lakka TA, Lind L, Dedoussis G, Njolstad I, Pedersen NL, Khaw KT, Wareham NJ, Keinanen-Kiukaanniemi SM, Saaristo TE, Korpi-Hyovalti E, Saltevo J, Laakso M, Kuusisto J, Metspalu A, Collins FS, Mohlke KL, Bergman RN, Tuomilehto J, Boehm BO, Gieger C, Hveem K, Cauchi S, Froguel P, Baldassarre D, Tremoli E, Humphries SE, Saleheen D, Danesh J, Ingelsson E, Ripatti S, Salomaa V, Erbel R, Jockel KH, Moebus S, Peters A, Illig T, de Faire U, Hamsten A, Morris AD, Donnelly PJ, Frayling TM, Hattersley AT, Boerwinkle E, Melander O, Kathiresan S, Nilsson PM, Deloukas P, Thorsteinsdottir U, Groop LC, Stefansson K, Hu F, Pankow JS, Dupuis J, Meigs JB, Altshuler D, Boehnke M, McCarthy MI, Replication DIG, Meta-analysis C. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nature genetics. 2012;44(9):981-90. doi: 10.1038/ng.2383. PubMed PMID: 22885922; PubMed Central PMCID: PMC3442244.
10. Replication DIG, Meta-analysis C, Asian Genetic Epidemiology Network Type 2 Diabetes C, South Asian Type 2 Diabetes C, Mexican American Type 2 Diabetes C, Type 2 Diabetes Genetic Exploration by Nex-generation sequencing in muylti-Ethnic Samples C, Mahajan A, Go MJ, Zhang W, Below JE, Gaulton KJ, Ferreira T, Horikoshi M, Johnson AD, Ng MC, Prokopenko I, Saleheen D, Wang X, Zeggini E, Abecasis GR, Adair LS, Almgren P, Atalay M, Aung T, Baldassarre D, Balkau B, Bao Y, Barnett AH, Barroso I, Basit A, Been LF, Beilby J, Bell GI, Benediktsson R, Bergman RN, Boehm BO, Boerwinkle E, Bonnycastle LL, Burtt N, Cai Q, Campbell H, Carey J, Cauchi S, Caulfield M, Chan JC, Chang LC, Chang TJ, Chang YC, Charpentier G, Chen CH, Chen H, Chen YT, Chia KS, Chidambaram M, Chines PS, Cho NH, Cho YM, Chuang LM, Collins FS, Cornelis MC, Couper DJ, Crenshaw AT, van Dam RM, Danesh J, Das D, de

Faire U, Dedoussis G, Deloukas P, Dimas AS, Dina C, Doney AS, Donnelly PJ, Dorkhan M, van Duijn C, Dupuis J, Edkins S, Elliott P, Emilsson V, Erbel R, Eriksson JG, Escobedo J, Esko T, Eury E, Florez JC, Fontanillas P, Forouhi NG, Forsen T, Fox C, Fraser RM, Frayling TM, Froguel P, Frossard P, Gao Y, Gertow K, Gieger C, Gigante B, Grallert H, Grant GB, Grrop LC, Groves CJ, Grundberg E, Guiducci C, Hamsten A, Han BG, Hara K, Hassanali N, Hattersley AT, Hayward C, Hedman AK, Herder C, Hofman A, Holmen OL, Hovingh K, Hreidarsson AB, Hu C, Hu FB, Hui J, Humphries SE, Hunt SE, Hunter DJ, Hveem K, Hydrie ZI, Ikegami H, Illig T, Ingelsson E, Islam M, Isomaa B, Jackson AU, Jafar T, James A, Jia W, Jockel KH, Jonsson A, Jowett JB, Kadowaki T, Kang HM, Kanoni S, Kao WH, Kathiresan S, Kato N, Katulanda P, Keinanen-Kiukaanniemi KM, Kelly AM, Khan H, Khaw KT, Khor CC, Kim HL, Kim S, Kim YJ, Kinnunen L, Klopp N, Kong A, Korpi-Hyovalti E, Kowlessur S, Kraft P, Kravic J, Kristensen MM, Krithika S, Kumar A, Kumate J, Kuusisto J, Kwak SH, Laakso M, Lagou V, Lakka TA, Langenberg C, Langford C, Lawrence R, Leander K, Lee JM, Lee NR, Li M, Li X, Li Y, Liang J, Liju S, Lim WY, Lind L, Lindgren CM, Lindholm E, Liu CT, Liu JJ, Lobbens S, Long J, Loos RJ, Lu W, Luan J, Lyssenko V, Ma RC, Maeda S, Magi R, Mannisto S, Matthews DR, Meigs JB, Melander O, Metspalu A, Meyer J, Mirza G, Mihailov E, Moebus S, Mohan V, Mohlke KL, Morris AD, Muhleisen TW, Muller-Nurasyid M, Musk B, Nakamura J, Nakashima E, Navarro P, Ng PK, Nica AC, Nilsson PM, Njolstad I, Nothen MM, Ohnaka K, Ong TH, Owen KR, Palmer CN, Pankow JS, Park KS, Parkin M, Pechlivanis S, Pedersen NL, Peltonen L, Perry JR, Peters A, Pinidiyapathirage JM, Platou CG, Potter S, Price JF, Qi L, Radha V, Rallidis L, Rasheed A, Rathman W, Rauramaa R, Raychaudhuri S, Rayner NW, Rees SD, Rehnberg E, Ripatti S, Robertson N, Roden M, Rossin EJ, Rudan I, Rybin D, Saaristo TE, Salomaa V, Saltevo J, Samuel M, Sanghera DK, Saramies J, Scott J, Scott LJ, Scott RA, Segre AV, Sehmi J, Sennblad B, Shah N, Shah S, Shera AS, Shu XO, Shuldiner AR, Sigurdsson G, Sijbrands E, Silveira A, Sim X, Sivapalaratnam S, Small KS, So WY, Stancakova A, Stefansson K, Steinbach G, Steinthorsdottir V, Stirrups K, Strawbridge RJ, Stringham HM, Sun Q, Suo C, Syvanen AC, Takayanagi R, Takeuchi F, Tay WT, Teslovich TM, Thorand B, Thorleifsson G, Thorsteinsdottir U, Tikkanen E, Trakalo J, Tremoli E, Trip MD, Tsai FJ, Tuomi T, Tuomilehto J, Uitterlinden AG, Valladares-Salgado A, Vedantam S, Veglia F, Voight BF, Wang C, Wareham NJ, Wennauer R, Wickremasinghe AR, Wilsgaard T, Wilson JF, Wiltshire S, Winckler W, Wong TY, Wood AR, Wu JY, Wu Y, Yamamoto K, Yamauchi T, Yang M, Yengo L, Yokota M, Young R, Zabaneh D, Zhang F, Zhang R, Zheng W, Zimmet PZ, Altshuler D, Bowden DW, Cho YS, Cox NJ, Cruz M, Hanis CL, Kooner J, Lee JY, Seielstad M, Teo YY, Boehnke M, Parra EJ, Chambers JC, Tai ES, McCarthy MI, Morris AP. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. Nature genetics. 2014;46(3):234-44. doi: 10.1038/ng.2897. PubMed PMID: 24509480; PubMed Central PMCID: PMC3969612.

11. Karlsson FH, Tremaroli V, Nookaew I, Bergstrom G, Behre CJ, Fagerberg B, Nielsen J, Backhed F. Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature. 2013;498(7452):99-103. doi: 10.1038/nature12198. PubMed PMID: 23719380.

12. de Vos WM, Nieuwdorp M. Genomics: A gut prediction. Nature. 2013;498(7452):48-9. doi: 10.1038/nature12251. PubMed PMID: 23719383.

13. Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, Liang S, Zhang W, Guan Y, Shen D, Peng Y, Zhang D, Jie Z, Wu W, Qin Y, Xue W, Li J, Han L, Lu D, Wu P, Dai Y, Sun X, Li Z, Tang A, Zhong S, Li X, Chen W, Xu R, Wang M, Feng Q, Gong M, Yu J, Zhang Y, Zhang M, Hansen T, Sanchez G, Raes J, Falony G, Okuda S, Almeida M, LeChatelier E, Renault P, Pons N, Batto JM, Zhang Z, Chen H, Yang R, Zheng W, Li S, Yang H, Wang J, Ehrlich SD, Nielsen R, Pedersen O, Kristiansen K, Wang J. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature. 2012;490(7418):55-60. Epub 2012/10/02. doi: 10.1038/nature11450. PubMed PMID: 23023125.

14. Oh J, Segre JA. Genomics: Resident risks. Nature. 2012;490(7418):44-6. doi: 10.1038/490044a. PubMed PMID: 23038462; PubMed Central PMCID: PMC3513833.

15. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP, Heath AC, Warner B, Reeder J, Kuczynski J, Caporaso JG, Lozupone CA, Lauber C, Clemente JC, Knights D, Knight R, Gordon JI. Human gut microbiome viewed across age and geography. Nature. 2012;486(7402):222-7. doi: 10.1038/nature11053. PubMed PMID: 22699611; PubMed Central PMCID: PMC3376388.

16. Turnbaugh PJ, Gordon JI. The core gut microbiome, energy balance and obesity. The Journal of physiology. 2009;587(Pt 17):4153-8. doi: 10.1113/jphysiol.2009.174136. PubMed PMID: 19491241; PubMed Central PMCID: PMC2754355.

17. Larsen N, Vogensen FK, van den Berg FW, Nielsen DS, Andreasen AS, Pedersen BK, Al-Soud WA, Sorensen SJ, Hansen LH, Jakobsen M. Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. PloS one. 2010;5(2):e9085. doi: 10.1371/journal.pone.0009085. PubMed PMID: 20140211; PubMed Central PMCID: PMC2816710.

18. Zhang X, Shen D, Fang Z, Jie Z, Qiu X, Zhang C, Chen Y, Ji L. Human gut microbiota changes reveal the progression of glucose intolerance. PloS one. 2013;8(8):e71108. doi: 10.1371/journal.pone.0071108. PubMed PMID: 24013136; PubMed Central PMCID: PMC3754967.

19. Grice EA, Segre JA. The human microbiome: our second genome. Annual review of genomics and human genetics. 2012;13:151-70. doi: 10.1146/annurev-genom-090711-163814. PubMed PMID: 22703178; PubMed Central PMCID: PMC3518434.
20. Dorrestein PC, Mazmanian SK, Knight R. Finding the missing links among metabolites, microbes, and the host. Immunity. 2014;40(6):824-32. Epub 2014/06/21. doi: 10.1016/j.immuni.2014.05.015. PubMed PMID: 24950202.
21. Goodrich JK, Di Rienzi SC, Poole AC, Koren O, Walters WA, Caporaso JG, Knight R, Ley RE. Conducting a Microbiome Study. Cell. 2014;158(2):250-62. Epub 2014/07/19. doi: 10.1016/j.cell.2014.06.037. PubMed PMID: 25036628.
22. Kootte RS, Vrieze A, Holleman F, Dallinga-Thie GM, Zoetendal EG, de Vos WM, Groen AK, Hoekstra JB, Stroes ES, Nieuwdorp M. The therapeutic potential of manipulating gut microbiota in obesity and type 2 diabetes mellitus. Diabetes, obesity & metabolism. 2012;14(2):112-20. Epub 2011/08/05. doi: 10.1111/j.1463-1326.2011.01483.x. PubMed PMID: 21812894.
23. Panwar H, Rashmi HM, Batish VK, Grover S. Probiotics as potential biotherapeutics in the management of type 2 diabetes - prospects and perspectives. Diabetes/metabolism research and reviews. 2013;29(2):103-12. doi: 10.1002/dmrr.2376. PubMed PMID: 23225499.
24. Vrieze A, Van Nood E, Holleman F, Salojarvi J, Kootte RS, Bartelsman JF, Dallinga-Thie GM, Ackermans MT, Serlie MJ, Oozeer R, Derrien M, Druesne A, Van Hylckama Vlieg JE, Bloks VW, Groen AK, Heilig HG, Zoetendal EG, Stroes ES, de Vos WM, Hoekstra JB, Nieuwdorp M. Transfer of intestinal microbiota from lean donors increases insulin sensitivity in individuals with metabolic syndrome. Gastroenterology. 2012;143(4):913-6 e7. doi: 10.1053/j.gastro.2012.06.031. PubMed PMID: 22728514.
25. Gonzalez A, Knight R. Advancing analytical algorithms and pipelines for billions of microbial sequences. Current opinion in biotechnology. 2012;23(1):64-71. Epub 2011/12/17. doi: 10.1016/j.copbio.2011.11.028. PubMed PMID: 22172529; PubMed Central PMCID: PMC3273654.
26. Fraher MH, O'Toole PW, Quigley EM. Techniques used to characterize the gut microbiota: a guide for the clinician. Nature reviews Gastroenterology & hepatology. 2012;9(6):312-22. doi: 10.1038/nrgastro.2012.44. PubMed PMID: 22450307.
27. Kau AL, Ahern PP, Griffin NW, Goodman AL, Gordon JI. Human nutrition, the gut microbiome and the immune system. Nature. 2011;474(7351):327-36. doi: 10.1038/nature10213. PubMed PMID: 21677749; PubMed Central PMCID: PMC3298082.
28. Everard A, Cani PD. Diabetes, obesity and gut microbiota. Best practice & research Clinical gastroenterology. 2013;27(1):73-83. doi: 10.1016/j.bpg.2013.03.007. PubMed PMID: 23768554.
29. Evans JM, Morris LS, Marchesi JR. The gut microbiome: the role of a virtual organ in the endocrinology of the host. The Journal of endocrinology. 2013;218(3):R37-47. doi: 10.1530/JOE-13-0131. PubMed PMID: 23833275.
30. Tilg H, Kaser A. Gut microbiome, obesity, and metabolic dysfunction. The Journal of clinical investigation. 2011;121(6):2126-32. doi: 10.1172/JCI58109. PubMed PMID: 21633181; PubMed Central PMCID: PMC3104783.
31. Greiner T, Backhed F. Effects of the gut microbiota on obesity and glucose homeostasis. Trends in endocrinology and metabolism: TEM. 2011;22(4):117-23. doi: 10.1016/j.tem.2011.01.002. PubMed PMID: 21353592.
32. Delzenne NM, Cani PD. Gut microbiota and the pathogenesis of insulin resistance. Current diabetes reports. 2011;11(3):154-9. doi: 10.1007/s11892-011-0191-1. PubMed PMID: 21431853.
33. Johnson AM, Olefsky JM. The origins and drivers of insulin resistance. Cell. 2013;152(4):673-84. doi: 10.1016/j.cell.2013.01.041. PubMed PMID: 23415219.
34. Roelofsen H, Priebe MG, Vonk RJ. The interaction of short-chain fatty acids with adipose tissue: relevance for prevention of type 2 diabetes. Beneficial microbes. 2010;1(4):433-7. doi: 10.3920/BM2010.0028. PubMed PMID: 21831781.
35. Ridaura VK, Faith JJ, Rey FE, Cheng J, Duncan AE, Kau AL, Griffin NW, Lombard V, Henrissat B, Bain JR, Muehlbauer MJ, Ilkayeva O, Semenkovich CF, Funai K, Hayashi DK, Lyle BJ, Martini MC, Ursell LK, Clemente JC, Van Treuren W, Walters WA, Knight R, Newgard CB, Heath AC, Gordon JI. Gut microbiota from twins discordant for obesity modulate metabolism in mice. Science. 2013;341(6150):1241214. doi: 10.1126/science.1241214. PubMed PMID: 24009397; PubMed Central PMCID: PMC3829625.
36. Membrez M, Blancher F, Jaquet M, Bibiloni R, Cani PD, Burcelin RG, Corthesy I, Mace K, Chou CJ. Gut microbiota modulation with norfloxacin and ampicillin enhances glucose tolerance in mice. FASEB journal : official publication of the Federation of American Societies for Experimental Biology. 2008;22(7):2416-26. doi: 10.1096/fj.07-102723. PubMed PMID: 18326786.
37. Carvalho BM, Guadagnini D, Tsukumo DM, Schenka AA, Latuf-Filho P, Vassallo J, Dias JC, Kubota LT, Carvalheira JB, Saad MJ. Modulation of gut microbiota by antibiotics improves insulin signalling in high-fat fed mice. Diabetologia. 2012;55(10):2823-34. doi: 10.1007/s00125-012-2648-4. PubMed PMID: 22828956.

38. Bech-Nielsen GV, Hansen CH, Hufeldt MR, Nielsen DS, Aasted B, Vogensen FK, Midtvedt T, Hansen AK. Manipulation of the gut microbiota in C57BL/6 mice changes glucose tolerance without affecting weight development and gut mucosal immunity. Research in veterinary science. 2012;92(3):501-8. doi: 10.1016/j.rvsc.2011.04.005. PubMed PMID: 21543097.
39. Human Microbiome Project C. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486(7402):207-14. doi: 10.1038/nature11234. PubMed PMID: 22699609; PubMed Central PMCID: PMC3564958.
40. 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. Proc Natl Acad Sci U S A. 2010;107(33):14691-6. doi: 10.1073/pnas.1005963107. PubMed PMID: 20679230; PubMed Central PMCID: PMCPMC2930426.
41. Blaut M, Clavel T. Metabolic diversity of the intestinal microbiota: implications for health and disease. J Nutr. 2007;137(3 Suppl 2):751S-5S. PubMed PMID: 17311972.
42. Backhed F, Ley RE, Sonnenburg JL, Peterson DA, Gordon JI. Host-bacterial mutualism in the human intestine. Science. 2005;307(5717):1915-20. doi: 10.1126/science.1104816. PubMed PMID: 15790844.
43. Hooper LV, Midtvedt T, Gordon JI. How host-microbial interactions shape the nutrient environment of the mammalian intestine. Annu Rev Nutr. 2002;22:283-307. doi: 10.1146/annurev.nutr.22.011602.092259. PubMed PMID: 12055347.
44. Glick-Bauer M, Yeh MC. The health advantage of a vegan diet: exploring the gut microbiota connection. Nutrients. 2014;6(11):4822-38. doi: 10.3390/nu6114822. PubMed PMID: 25365383; PubMed Central PMCID: PMCPMC4245565.
45. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, Ling AV, Devlin AS, Varma Y, Fischbach MA, Biddinger SB, Dutton RJ, Turnbaugh PJ. Diet rapidly and reproducibly alters the human gut microbiome. Nature. 2014;505(7484):559-63. Epub 2013/12/18. doi: 10.1038/nature12820. PubMed PMID: 24336217; PubMed Central PMCID: PMC3957428.
46. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen YY, Keilbaugh SA, Bewtra M, Knights D, Walters WA, Knight R, Sinha R, Gilroy E, Gupta K, Baldassano R, Nessel L, Li H, Bushman FD, Lewis JD. Linking long-term dietary patterns with gut microbial enterotypes. Science. 2011;334(6052):105-8. Epub 2011/09/03. doi: 10.1126/science.1208344. PubMed PMID: 21885731; PubMed Central PMCID: PMC3368382.
47. Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, Raes J, Huttenhower C. Microbial co-occurrence relationships in the human microbiome. PLoS computational biology. 2012;8(7):e1002606. Epub 2012/07/19. doi: 10.1371/journal.pcbi.1002606. PubMed PMID: 22807668; PubMed Central PMCID: PMCPmc3395616.
48. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto JM, Bertalan M, Borruel N, Casellas F, Fernandez L, Gautier L, Hansen T, Hattori M, Hayashi T, Kleerebezem M, Kurokawa K, Leclerc M, Levenez F, Manichanh C, Nielsen HB, Nielsen T, Pons N, Poulain J, Qin J, Sicheritz-Ponten T, Tims S, Torrents D, Ugarte E, Zoetendal EG, Wang J, Guarner F, Pedersen O, de Vos WM, Brunak S, Dore J, Meta HITC, Antolin M, Artiguenave F, Blottiere HM, Almeida M, Brechot C, Cara C, Chervaux C, Cultrone A, Delorme C, Denariaz G, Dervyn R, Foerstner KU, Friss C, van de Guchte M, Guedon E, Haimet F, Huber W, van Hylckama-Vlieg J, Jamet A, Juste C, Kaci G, Knol J, Lakhdari O, Layec S, Le Roux K, Maguin E, Merieux A, Melo Minardi R, M'Rini C, Muller J, Oozeer R, Parkhill J, Renault P, Rescigno M, Sanchez N, Sunagawa S, Torrejon A, Turner K, Vandemeulebrouck G, Varela E, Winogradsky Y, Zeller G, Weissenbach J, Ehrlich SD, Bork P. Enterotypes of the human gut microbiome. Nature. 2011;473(7346):174-80. Epub 2011/04/22. doi: 10.1038/nature09944. PubMed PMID: 21508958; PubMed Central PMCID: PMC3728647.
49. Canani RB, Costanzo MD, Leone L, Pedata M, Meli R, Calignano A. Potential beneficial effects of butyrate in intestinal and extraintestinal diseases. World J Gastroenterol. 2011;17(12):1519-28. doi: 10.3748/wjg.v17.i12. 1519. PubMed PMID: 21472114; PubMed Central PMCID: PMCPMC3070119.
50. Kovatcheva-Datchary P, Nilsson A, Akrami R, Lee YS, De Vadder F, Arora T, Hallen A, Martens E, Bjorck I, Backhed F. Dietary Fiber-Induced Improvement in Glucose Metabolism Is Associated with Increased Abundance of Prevotella. Cell Metab. 2015;22(6):971-82. doi: 10.1016/j.cmet.2015.10.001. PubMed PMID: 26552345.
51. Fredricks DN, Relman DA. Infectious agents and the etiology of chronic idiopathic diseases. Current clinical topics in infectious diseases. 1998;18:180-200. PubMed PMID: 9779355.
52. Angelakis E, Merhej V, Raoult D. Related actions of probiotics and antibiotics on gut microbiota and weight modification. The Lancet infectious diseases. 2013;13(10):889-99. doi: 10.1016/S1473-3099(13)70179-8. PubMed PMID: 24070562.
53. Aggarwal J, Swami G, Kumar M. Probiotics and their Effects on Metabolic Diseases: An Update. Journal of clinical and diagnostic research : JCDR. 2013;7(1):173-7. doi: 10.7860/JCDR/2012/5004.2701. PubMed PMID: 23449881; PubMed Central PMCID: PMC3576782.
54. Lemon KP, Armitage GC, Relman DA, Fischbach MA. Microbiota-targeted therapies: an ecological perspective. Science translational medicine. 2012;4(137):137rv5. doi: 10.1126/scitranslmed.3004183. PubMed PMID: 22674555.

55. Faith JJ, Guruge JL, Charbonneau M, Subramanian S, Seedorf H, Goodman AL, Clemente JC, Knight R, Heath AC, Leibel RL, Rosenbaum M, Gordon JI. The long-term stability of the human gut microbiota. Science. 2013;341(6141):1237439. doi: 10.1126/science.1237439. PubMed PMID: 23828941; PubMed Central PMCID: PMC3791589.
56. Rajilic-Stojanovic M, Heilig HG, Tims S, Zoetendal EG, de Vos WM. Long-term monitoring of the human intestinal microbiota composition. Environmental microbiology. 2012. Epub 2013/01/05. doi: 10.1111/1462-2920.12023. PubMed PMID: 23286720.
57. Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R. Diversity, stability and resilience of the human gut microbiota. Nature. 2012;489(7415):220-30. Epub 2012/09/14. doi: 10.1038/nature11550. PubMed PMID: 22972295; PubMed Central PMCID: PMC3577372.
58. Carroll IM, Ringel-Kulka T, Siddle JP, Klaenhammer TR, Ringel Y. Characterization of the fecal microbiota using high-throughput sequencing reveals a stable microbial community during storage. PloS one. 2012;7(10):e46953. Epub 2012/10/17. doi: 10.1371/journal.pone.0046953. PubMed PMID: 23071673; PubMed Central PMCID: PMC3465312.
59. Relman DA. The human microbiome: ecosystem resilience and health. Nutrition reviews. 2012;70 Suppl 1:S2-9. doi: 10.1111/j.1753-4887.2012.00489.x. PubMed PMID: 22861804; PubMed Central PMCID: PMC3422777.
60. Griffin JL, Wang X, Stanley E. Does our gut microbiome predict cardiovascular risk? A review of the evidence from metabolomics. Circulation Cardiovascular genetics. 2015;8(1):187-91. doi: 10.1161/CIRCGENETICS.114.000219. PubMed PMID: 25691688; PubMed Central PMCID: PMC4333723.
61. Lavange LM, Kalsbeek WD, Sorlie PD, Aviles-Santa LM, Kaplan RC, Barnhart J, Liu K, Giachello A, Lee DJ, Ryan J, Criqui MH, Elder JP. Sample design and cohort selection in the Hispanic Community Health Study/Study of Latinos. Annals of epidemiology. 2010;20(8):642-9. doi: 10.1016/j.annepidem.2010.05.006. PubMed PMID: 20609344; PubMed Central PMCID: PMC2921622.
62. Sorlie PD, Aviles-Santa LM, Wassertheil-Smoller S, Kaplan RC, Daviglus ML, Giachello AL, Schneiderman N, Raij L, Talavera G, Allison M, Lavange L, Chambless LE, Heiss G. Design and implementation of the Hispanic Community Health Study/Study of Latinos. Annals of epidemiology. 2010;20(8):629-41. doi: 10.1016/j.annepidem.2010.03.015. PubMed PMID: 20609343; PubMed Central PMCID: PMC2904957.
63. Vazquez-Baeza Y, Pirrung M, Gonzalez A, Knight R. EMPeror: a tool for visualizing high-throughput microbial community data. GigaScience. 2013;2(1):16. Epub 2013/11/28. doi: 10.1186/2047-217X-2-16. PubMed PMID: 24280061; PubMed Central PMCID: PMC4076506.
64. Navas-Molina JA, Peralta-Sanchez JM, Gonzalez A, McMurdie PJ, Vazquez-Baeza Y, Xu Z, Ursell LK, Lauber C, Zhou H, Song SJ, Huntley J, Ackermann GL, Berg-Lyons D, Holmes S, Caporaso JG, Knight R. Advancing our understanding of the human microbiome using QIIME. Methods in enzymology. 2013;531:371-444. Epub 2013/09/26. doi: 10.1016/B978-0-12-407863-5.00019-8. PubMed PMID: 24060131.
65. Brown J, de Vos WM, DiStefano PS, Dore J, Huttenhower C, Knight R, Lawley TD, Raes J, Turnbaugh P. Translating the human microbiome. Nature biotechnology. 2013;31(4):304-8. doi: 10.1038/nbt.2543. PubMed PMID: 23563424.
66. Martinez I, Muller CE, Walter J. Long-term temporal analysis of the human fecal microbiota revealed a stable core of dominant bacterial species. PloS one. 2013;8(7):e69621. doi: 10.1371/journal.pone.0069621. PubMed PMID: 23874976; PubMed Central PMCID: PMC3712949.
67. Flores R, Shi J, Gail MH, Gajer P, Ravel J, Goedert JJ. Assessment of the human faecal microbiota: II. Reproducibility and associations of 16S rRNA pyrosequences. European journal of clinical investigation. 2012;42(8):855-63. Epub 2012/03/06. doi: 10.1111/j.1365-2362.2012.02659.x. PubMed PMID: 22385292; PubMed Central PMCID: PMC3369017.
68. Lauber CL, Zhou N, Gordon JI, Knight R, Fierer N. Effect of storage conditions on the assessment of bacterial community structure in soil and human-associated samples. FEMS microbiology letters. 2010;307(1):80-6. Epub 2010/04/24. doi: 10.1111/j.1574-6968.2010.01965.x. PubMed PMID: 20412303; PubMed Central PMCID: PMC3148093.
69. American Diabetes A. Diagnosis and classification of diabetes mellitus. Diabetes care. 2013;36 Suppl 1:S67-74. doi: 10.2337/dc13-S067. PubMed PMID: 23264425; PubMed Central PMCID: PMC3537273.
70. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M, Gormley N, Gilbert JA, Smith G, Knight R. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J. 2012;6(8):1621-4. doi: 10.1038/ismej.2012.8. PubMed PMID: 22402401; PubMed Central PMCID: PMC3400413.
71. Magoc T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics. 2011;27(21):2957-63. doi: 10.1093/bioinformatics/btr507. PubMed PMID: 21903629; PubMed Central PMCID: PMC3198573.

72. 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. Nat Methods. 2013;10(1):57-9. doi: 10.1038/nmeth.2276. PubMed PMID: 23202435; PubMed Central PMCID: PMC3531572.
73. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. 2012;6(3):610-8. doi: 10.1038/ismej.2011.139. PubMed PMID: 22134646; PubMed Central PMCID: PMC3280142.
74. 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(7):5069-72. doi: 10.1128/AEM.03006-05. PubMed PMID: 16820507; PubMed Central PMCID: PMC1489311.
75. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013;41(Database issue):D590-6. doi: 10.1093/nar/gks1219. PubMed PMID: 23193283; PubMed Central PMCID: PMC3531112.
76. Caporaso JG, Bittinger K, Bushman FD, DeSantis TZ, Andersen GL, Knight R. PyNAST: a flexible tool for aligning sequences to a template alignment. Bioinformatics. 2010;26(2):266-7. doi: 10.1093/bioinformatics/btp636. PubMed PMID: 19914921; PubMed Central PMCID: PMC2804299.
77. Gevers D, Knight R, Petrosino JF, Huang K, McGuire AL, Birren BW, Nelson KE, White O, Methe BA, Huttenhower C. The Human Microbiome Project: a community resource for the healthy human microbiome. PLoS biology. 2012;10(8):e1001377. doi: 10.1371/journal.pbio.1001377. PubMed PMID: 22904687; PubMed Central PMCID: PMC3419203.
78. Human Microbiome Project C. A framework for human microbiome research. Nature. 2012;486(7402):215-21. doi: 10.1038/nature11209. PubMed PMID: 22699610; PubMed Central PMCID: PMC3377744.
79. Gilbert JA, Meyer F, Antonopoulos D, Balaji P, Brown CT, Brown CT, Desai N, Eisen JA, Evers D, Field D, Feng W, Huson D, Jansson J, Knight R, Knight J, Kolker E, Konstantindis K, Kostka J, Kyrpides N, Mackelprang R, McHardy A, Quince C, Raes J, Sczyrba A, Shade A, Stevens R. Meeting report: the terabase metagenomics workshop and the vision of an Earth microbiome project. Standards in genomic sciences. 2010;3(3):243-8. doi: 10.4056/sigs.1433550. PubMed PMID: 21304727; PubMed Central PMCID: PMC3035311.
80. Ley RE, Backhed F, Turnbaugh P, Lozupone CA, Knight RD, Gordon JI. Obesity alters gut microbial ecology. Proceedings of the National Academy of Sciences of the United States of America. 2005;102(31):11070-5. doi: 10.1073/pnas.0504978102. PubMed PMID: 16033867; PubMed Central PMCID: PMC1176910.
81. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, Egholm M, Henrissat B, Heath AC, Knight R, Gordon JI. A core gut microbiome in obese and lean twins. Nature. 2009;457(7228):480-4. doi: 10.1038/nature07540. PubMed PMID: 19043404; PubMed Central PMCID: PMC2677729.
82. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI. An obesity-associated gut microbiome with increased capacity for energy harvest. Nature. 2006;444(7122):1027-31. doi: 10.1038/nature05414. PubMed PMID: 17183312.
83. Liu Z, DeSantis TZ, Andersen GL, Knight R. Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. Nucleic Acids Res. 2008;36(18):e120. doi: 10.1093/nar/gkn491. PubMed PMID: 18723574; PubMed Central PMCID: PMC2566877.
84. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. Applied and environmental microbiology. 2005;71(12):8228-35. doi: 10.1128/AEM.71.12.8228-8235.2005. PubMed PMID: 16332807; PubMed Central PMCID: PMC1317376.
85. Lozupone C, Lladser ME, Knights D, Stombaugh J, Knight R. UniFrac: an effective distance metric for microbial community comparison. ISME J. 2011;5(2):169-72. doi: 10.1038/ismej.2010.133. PubMed PMID: 20827291; PubMed Central PMCID: PMC3105689.
86. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. 2014;42(Database issue):D199-205. doi: 10.1093/nar/gkt1076. PubMed PMID: 24214961; PubMed Central PMCID: PMC3965122.
87. Smillie CS, Smith MB, Friedman J, Cordero OX, David LA, Alm EJ. Ecology drives a global network of gene exchange connecting the human microbiome. Nature. 2011;480(7376):241-4. doi: 10.1038/nature10571. PubMed PMID: 22037308.
88. Zaneveld JR, Lozupone C, Gordon JI, Knight R. Ribosomal RNA diversity predicts genome diversity in gut bacteria and their relatives. Nucleic Acids Res. 2010;38(12):3869-79. doi: 10.1093/nar/gkq066. PubMed PMID: 20197316; PubMed Central PMCID: PMC2896507.

89. David LA, Materna AC, Friedman J, Campos-Baptista MI, Blackburn MC, Perrotta A, Erdman SE, Alm EJ. Host lifestyle affects human microbiota on daily timescales. Genome biology. 2014;15(7):R89. Epub 2014/08/26. doi: 10.1186/gb-2014-15-7-r89. PubMed PMID: 25146375.
90. Dassi E, Ballarini A, Covello G, Htm CMB, Quattrone A, Jousson O, De Sanctis V, Bertorelli R, Denti MA, Segata N. Enhanced microbial diversity in the saliva microbiome induced by short-term probiotic intake revealed by 16S rRNA sequencing on the IonTorrent PGM platform. Journal of biotechnology. 2014;190:30-9. doi: 10.1016/j.jbiotec.2014.03.024. PubMed PMID: 24670254.
91. Abubucker S, Segata N, Goll J, Schubert AM, Izard J, Cantarel BL, Rodriguez-Mueller B, Zucker J, Thiagarajan M, Henrissat B, White O, Kelley ST, Methe B, Schloss PD, Gevers D, Mitreva M, Huttenhower C. Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS computational biology. 2012;8(6):e1002358. doi: 10.1371/journal.pcbi.1002358. PubMed PMID: 22719234; PubMed Central PMCID: PMC3374609.
92. Podell S, Gaasterland T. DarkHorse: a method for genome-wide prediction of horizontal gene transfer. Genome biology. 2007;8(2):R16. doi: 10.1186/gb-2007-8-2-r16. PubMed PMID: 17274820; PubMed Central PMCID: PMC1852411.
93. McDonald D, Clemente JC, Kuczynski J, Rideout JR, Stombaugh J, Wendel D, Wilke A, Huse S, Hufnagle J, Meyer F, Knight R, Caporaso JG. The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. GigaScience. 2012;1(1):7. doi: 10.1186/2047-217X-1-7. PubMed PMID: 23587224; PubMed Central PMCID: PMC3626512.
94. Mandal S, Van Treuren W, White RA, Eggesbo M, Knight R, Peddada SD. Analysis of composition of microbiomes: a novel method for studying microbial composition. Microbial ecology in health and disease. 2015;26:27663. doi: 10.3402/mehd.v26.27663. PubMed PMID: 26028277; PubMed Central PMCID: PMC4450248.
95. McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PloS one. 2013;8(4):e61217. doi: 10.1371/journal.pone.0061217. PubMed PMID: 23630581; PubMed Central PMCID: PMC3632530.
96. Rubin D. Multiple imputation for nonresponse in surveys. New York: J. Wiley & Sons; 1987.
97. Schafer JL. Analysis of incomplete multivariate data. London: Chapman & Hall/CRC Press; 1997.
98. Thornton T, Tang H, Hoffmann TJ, Ochs-Balcom HM, Caan BJ, Risch N. Estimating kinship in admixed populations. American journal of human genetics. 2012;91(1):122-38. doi: 10.1016/j.ajhg.2012.05.024. PubMed PMID: 22748210; PubMed Central PMCID: PMC3397261.
99. Koren O, Knights D, Gonzalez A, Waldron L, Segata N, Knight R, Huttenhower C, Ley RE. A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. PLoS computational biology. 2013;9(1):e1002863. doi: 10.1371/journal.pcbi.1002863. PubMed PMID: 23326225; PubMed Central PMCID: PMC3542080.
100. Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, Lander ES, Mitzenmacher M, Sabeti PC. Detecting novel associations in large data sets. Science. 2011;334(6062):1518-24. doi: 10.1126/science.1205438. PubMed PMID: 22174245; PubMed Central PMCID: PMC3325791.
101. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological). 1995:289-300.
102. Chen J, Li H. Variable Selection for Sparse Dirichlet-Multinomial Regression with an Application to Microbiome Data Analysis. Ann Appl Stat. 2013;7(1):418-42. doi: 10.1214/12-AOAS592. PubMed PMID: 24312162; PubMed Central PMCID: PMC3846354.
103. Manichaikul A, Palmas W, Rodriguez CJ, Peralta CA, Divers J, Guo X, Chen WM, Wong Q, Williams K, Kerr KF, Taylor KD, Tsai MY, Goodarzi MO, Sale MM, Diez-Roux AV, Rich SS, Rotter JI, Mychaleckyj JC. Population structure of Hispanics in the United States: the multi-ethnic study of atherosclerosis. PLoS genetics. 2012;8(4):e1002640. doi: 10.1371/journal.pgen.1002640. PubMed PMID: 22511882; PubMed Central PMCID: PMC3325201.
104. Conomos MP, Laurie CA, Stilp AM, Gogarten SM, McHugh CP, Nelson SC, Sofer T, Fernandez-Rhodes L, Justice AE, Graff M, Young KL, Seyerle AA, Avery CL, Taylor KD, Rotter JI, Talavera GA, Daviglus ML, Wassertheil-Smoller S, Schneiderman N, Heiss G, Kaplan RC, Franceschini N, Reiner AP, Shaffer JR, Barr RG, Kerr KF, Browning SR, Browning BL, Weir BS, Aviles-Santa ML, Papanicolaou GJ, Lumley T, Szpiro AA, North KE, Rice K, Thornton TA, Laurie CC. Genetic Diversity and Association Studies in US Hispanic/Latino Populations: Applications in the Hispanic Community Health Study/Study of Latinos. American journal of human genetics. 2016;98(1):165-84. doi: 10.1016/j.ajhg.2015.12.001. PubMed PMID: 26748518; PubMed Central PMCID: PMC4716704.
105. Browning SR, Grinde K, Plantinga A, Gogarten SM, Stilp AM, Kaplan RC, Aviles-Santa ML, Browning BL, Laurie C. Local ancestry inference in a large US-based Hispanic/Latino study: Hispanic Community Health Study / Study of Latinos (HCHS/SOL) G3: Genes, Genomes, Genetics. In Press.
106. Chen J, Bittinger K, Charlson ES, Hoffmann C, Lewis J, Wu GD, Collman RG, Bushman FD, Li H. Associating microbiome composition with environmental covariates using generalized UniFrac distances. Bioinformatics.

2012;28(16):2106-13. doi: 10.1093/bioinformatics/bts342. PubMed PMID: 22711789; PubMed Central PMCID: PMC3413390.
107. White JR, Nagarajan N, Pop M. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. PLoS Comput Biol. 2009;5(4):e1000352. Epub 2009/04/11. doi: 10.1371/journal.pcbi.1000352. PubMed PMID: 19360128; PubMed Central PMCID: PMC2661018.